DOCTOR OF PHILOSOPHY

Longitudinal Joint Modelling Utilising the Coxian Phase-type Distribution

Donnelly, Conor

Award date:
2019

Awarding institution:
Queen's University Belfast

Longitudinal Joint Modelling Utilising the Coxian Phase-type Distribution

Conor Donnelly, MSci (Hons)

Thesis submitted for the degree of

Doctor of Philosophy

in the

Faculty of Engineering and Physical Sciences

Queen's University Belfast

September 2018

Abstract

Joint modelling techniques for the analysis of longitudinal and survival data are a relatively recent statistical development, employed to appropriately account for the possible association which can exist between these two processes. Since their development at the end of the 20th century, joint models have become increasingly popular within statistical literature due to their broad applicability across many areas of statistical research. Within a joint modelling framework, each process is represented by its own submodel, where most often a linear mixed effects (LME) model is employed for the longitudinal process and a proportional hazards model for the survival outcome. The parameters of both submodels are estimated simultaneously from a single joint likelihood, therefore overcoming the bias which can occur when the processes are modelled independently.

Within this thesis, the Coxian phase-type regression model is explored as a novel approach to represent the survival process within a joint model. Phase-type distributions, in general, are a diverse family of distributions which describe the absorption times of a continuous time Markov process with a single absorbing state, formulated by a convolution of exponential distributions, either in series or parallel. Employing phase-type distributions to model failure times can potentially uncover latent stages of the process under investigation, and insight can be gained from the estimated parameters regarding the rates of flow through these uncovered phases. Within a medical context, mapping the uncovered states onto distinct stages of a disease's progression, for example, allows predictions to be made regarding the time spent within the different stages of the disease, and inferences can be drawn from these predictions on the individuals' expected quality of life for their remaining survival time.

Whilst previous research has explored the use of phase-type distributions to represent typical survival analysis problems, there are a number of limitations which have hindered their popularity, particularly with regards to the fitting of the models. Consequently, the first portion of this research is concerned with investigating the Coxian phase-type regression model so as to improve its suitability at representing typical survival processes. To this end, a new expectation-maximisation (EM) algorithm approach to fitting phase-type distributions is developed, shown through a simulation study to improve upon alternative algorithm approaches, employed within the current literature, in terms of both the accuracy of the parameter estimates and the rate of convergence. Due to its increased stability, this approach is then extended to allow for the effect of a covariate to vary across the transitions of the model, as opposed to remaining fixed, as is the common assumption within the literature.

Subsequently, a joint modelling framework is developed, within which the longitudinal process is represented by a LME model and the survival process by the Coxian phase-type regression model. This new methodology is shown to be beneficial to both areas of research. For instance, previously the Coxian was limited to modelling time-invariant covariates, greatly limiting its scope, whereas it can now be employed within the new joint modelling framework to model the association between longitudinal biomarkers and survival outcome, extending its applicability, particularly within the medical field.

Further, employing the Coxian to represent the survival process offers a number of advantages over alternative parametric models; the Coxian phase-type distribution can represent any positive distribution to an arbitrary degree of accuracy, overcoming the noted limitations of survival models which assume more restrictive distributions, and greater insight into the survival process can be gained from the uncovered phases of the distribution.

Acknowledgements

Firstly, I would like to thank my supervisors Dr Lisa McFetridge and Professor Adele Marshall, whose wisdom and encouragement has been invaluable to me throughout this PhD. I have thoroughly enjoyed working together over these past four years, and am so grateful for their support.

Secondly, I would like to thank my fellow PhD students in Lanyon South: Laura Boyle, Meabh McCurdy and Caoimhe Carbery. When I look back on this time I won't think of the stress, but rather the fun that was had and laughs that were shared. The greatest reward from this PhD is the lifelong friendships that were formed.

Finally I want to thank my friends and family who kept me sane throughout and supported me no matter what.

Communication of Work

Peer-reviewed Publications

Donnelly, C. & McFetridge, L. M. & Marshall A. H. & Mitchell, H. J. (2017) *A two-stage approach to the joint analysis of longitudinal and survival data utilising the Coxian phase-type distribution*, Statistical Methods in Medical Research, (in press)

Conference Presentations

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2018) *A joint model for the analysis of longitudinal and survival data utilising the Coxian phase-type distribution*, Annual Conference on Applied Statistics in Ireland (CASI), The Irish Statistical Association, 16-18 May 2018, Galway, Ireland

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2017) *Joint modelling of longitudinal and time-to-event data utilising the Coxian phase-type distribution*, The Statistical Analysis of Multi-Outcome Data (SAM), 3-4 July 2017, Liverpool, UK

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2017) *Incorporating covariate effects within phase-type distributions using an EM algorithm approach*, Annual Conference on Applied Statistics in Ireland (CASI), The Irish Statistical Association, 15-17 May 2017, Mullingar, Ireland

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2016) *Exploring the Coxian phase-type distribution within a joint model setting*, Population-based Time-to-event Analyses International Conference, London School of Hygiene and Tropical Medicine, 31 Aug -2 Sept 2016, London, UK

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2016) *Utilising the Coxian phase-type distribution to represent patient survival within a joint modelling framework*, Annual Conference of the International Society of Clinical Biostatistics (ISCB),

21-25 Aug 2016, Birmingham, UK

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2016) *A multivariate joint modelling approach to incorporate individuals' longitudinal response trajectories within the Coxian phase-type distribution*, Annual Conference on Applied Statistics in Ireland (CASI), The Irish Statistical Association, 16-18 May 2016, Limerick, Ireland

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2016) *A multivariate two-stage joint model utilising the Coxian phase-type distribution to represent the survival process*, Joint Modelling and Beyond, Universiteit Hasselt, 14-15 April 2016, Hasselt, Belgium

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2015) *A longitudinal model with individual repeated measures as predictors of a Coxian phase-type survival distribution*, International Conference of the ERCIM on Computational and Methodological Statistics (CMStatistics), University of London, 12-14 Dec 2015, London, UK

Donnelly, C. & McFetridge, L. M. & Marshall, A. H. (2015) *Comparison of Approaches to the Joint Analysis of Longitudinal and Survival Renal Data*, Annual Conference on Applied Statistics in Ireland (CASI), The Irish Statistical Association, 11-13 May 2015, Cork, Ireland

Contents

Chapter 1

Introduction

1.1 Overview

Both the fields of longitudinal and survival analysis have undergone significant statistical advancements within the last 50 years. Key methodological developments, such as the linear mixed effects (LME) model [1] for the analysis of unbalanced repeated measures data, and the Cox proportional hazards (PH) model [2] for the analysis of time-to-event processes, have had a profound impact upon these fields of research, within which they have come to dominate. By the end of the 20th century, a greater emphasis could be observed within the literature on the association which can exist between these two processes, culminating in the development of joint modelling techniques [3, 4]. Within a joint modelling framework, the longitudinal and survival processes are estimated simultaneously, through a single joint likelihood, allowing the effect of each process on the other to be appropriately taken into consideration, removing the bias which is well noted to contaminate parameter estimation within independent models [5, 6]. Despite their ever-increasing popularity, there still exists many areas of joint modelling which have not yet been fully explored, with much scope to both improve upon current estimation techniques and to investigate alternative representations of the two processes within the single joint likelihood.

This research explores the use of the Coxian phase-type regression model as a novel approach to represent typical survival data, initially as an independent model and subsequently within a joint modelling framework, offering a number of advantages when compared to standard survival representations. Namely, as phase-type distributions assume an underlying Markov process, applying such models to survival data can uncover latent states, or 'phases', of the failure process, and inferences can be made from the estimated parameters of the distribution on the rates of flow of individuals through these phases. This can be extremely useful within medical statistics, for instance, where the uncovered phases can be mapped onto distinct stages of the survival process, meaning greater insight can be gained regarding how individuals behave before experiencing their event of interest. For example, for chronic and degenerative conditions, where the uncovered states can represent increasingly severe stages of the disease, the models can provide insight into the quality of life an individual will experience during the remainder of their survival time by estimating how long they will spend within each state. Such information can be utilised by clinicians to inform treatment plans and target interventions towards patients when they are most in need and when the intervention will yield optimal results.

Incorporating such models within a joint framework, as is detailed within this research, significantly advances the theory of phase-type regression models, extending

their applicability to scenarios where time-varying covariates are of interest, not previously explored within the literature. Further, the field of joint modelling is similarly advanced, benefiting from additional features of phase-type distributions which overcome some well documented limitations of alternative distributional representations of the survival process [7]. Consequently, the research presented within this thesis contributes significantly to both the research areas of phase-type distributions and joint modelling.

A more detailed background of joint models and phase-type distributions is presented within Section 1.2, before a motivating example is briefly introduced within Section 1.3. Finally, the contributions of this research are detailed within Section 1.4, and an outline of the remainder of the thesis is given in Section 1.5.

1.2 Background

1.2.1 Joint Modelling of Longitudinal and Survival Data

The joint analysis of longitudinal and survival data is a relatively recent statistical technique, developed to take into consideration the association which often exists between these two processes. For instance, within survival analysis, it is often of interest to incorporate some dynamic, endogenous biomarker as a predictor of failure, where standard independent PH and acceleration failure time (AFT) models are ill-equipped to appropriately handle the time-varying nature of such covariates. Conversely, dropout from a longitudinal study due to an associated event process can result in underrepresentation of certain subpopulations, introducing bias to the estimated population-level parameters.

By 1995, it had been well established within both longitudinal and survival literature that the independent analysis of either process, in the presence of such an association, results in biased inferences [5, 6, 8]. Shortly thereafter, pioneering papers by Faucett and Thomas [3] in 1996, Wulfsohn and Tsiatis [4] in 1997, and Henderson et al. [9] in 2000, outlined new joint modelling methodology, where the two process are estimated simultaneously through a single joint likelihood. Within this initial modelling framework, the longitudinal process was represented by a LME model and the survival process by a Cox PH model, where the association between the two processes was represented by latent random effects.

Since the publication of these critical methodological developments, the area of joint modelling has grown rapidly, as illustrated by Figure 1.1, which shows an increasing trend in the number of citations for two key publications within the research

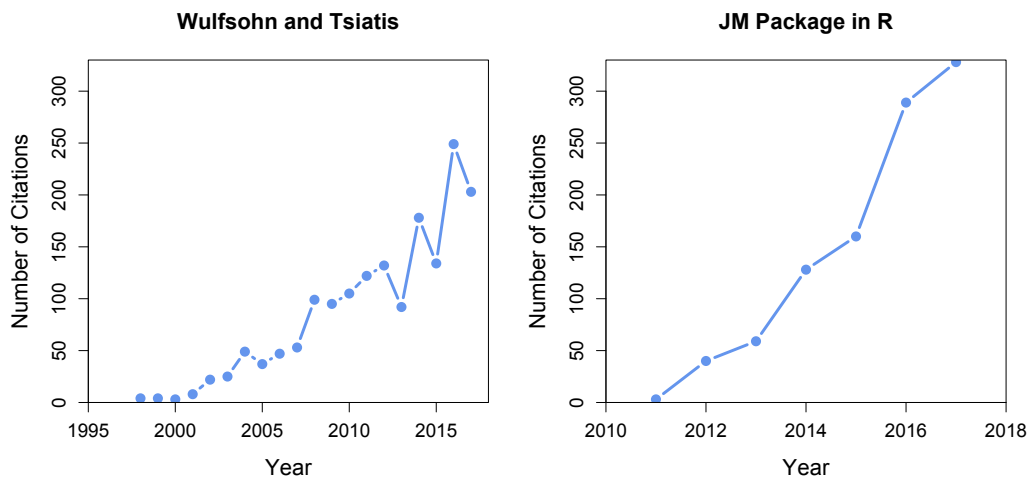


Figure 1.1: Graph showing the increasing number of citations for Wulfsohn and Tsiatis's pioneering paper [4] and the JM package within R software [14]. Data obtained from Web of Science, accessed September 2018 [15].

area. As such, many extensions to the methodology have been explored within the literature. For example, parametric AFT models have been incorporated within the joint modelling framework to represent the survival process [10], applicable in scenarios where the PH assumption does not hold. Similarly, fully parametric PH models have also been explored to represent the survival process, assuming, for example, an underlying exponential or Weibull distribution. Such fully parametric representations of the survival process, whilst successfully overcoming issues in estimating the standard errors which are experienced when employing the semi-parametric Cox PH model [11], are not without their limitations. As noted by Gould et al. [12], the exponential and Weibull distributions are limited in terms of the hazard shapes which they can represent, introducing error to the parameter estimates if the data does not truly observe the assumed distribution. This issue constitutes one of the targets within this research, where the Coxian phase-type distribution is advantageous due to its ability to represent any positive distribution to an arbitrary degree of accuracy [13], overcoming this well documented limitation of alternative survival representations within a joint framework.

Despite the conceptual simplicity of joint models, the fitting procedure involved to estimate the unknown parameters of the two processes can be computationally intensive, primarily due to the numerical approximations required within the likelihood. Indeed, as noted by Rizopoulos [16], these numerical approximations "constitute the

main reason why joint models have not yet found their rightful place in the toolbox of modern applied statisticians." More recent contributions to the field of joint modelling have, however, investigated approaches to improve the numerical approximations within the likelihood, where the standard Gauss-Hermite quadrature has been replaced with a pseudo-adaptive Gauss-Hermite approach [17], and the use of Laplace approximations to estimate these integrals has also been investigated [16].

Perhaps partially due to these computational difficulties, as well as their being a relatively recent statistical development, there exists only a limited number of approaches to fit joint models within modern statistical software, the most popular of which is the 'JM' package [14] within R [18]. First published in 2008, this package is designed to model the 'true longitudinal response' (TLR) parameterisation of the joint likelihood, where interest lies in the direct relationship between the true value of the longitudinal response and survival outcome. Within the package, the longitudinal process is represented by a LME model, and the survival process can be represented by a selection of submodels, such as: (i) a Cox PH model, (ii) an exponential or Weibull PH model, (iii) an exponential or Weibull AFT model, (iv) a PH model with a piecewise-constant baseline risk function, and (v) a PH model in which the log cumulative baseline hazard is approximated using B-splines. The piecewise-constant baseline formulation, along with the spline based approach, are suggested within the literature as suitable methods to overcome the limited distributional shapes which can be represented by the fully parametric PH models [19], however no equivalent AFT approaches are currently available within the software, limiting the choice to only an exponential or Weibull AFT model. Within this research, an AFT formulation of the Coxian phase-type regression model is incorporated within a joint framework, fulfilling the need for alternative representations of the baseline survival distribution when the Weibull does not suitably fit the shape of the data. This overcomes potential errors in the estimation of the survival parameters which can result from misspecifying the underlying distribution.

The TLR parameterisation can also be fitted using a Bayesian approach, by the R package 'JMBayes' [20]. However, the survival process can only be represented, thus far, by a proportional hazards model where the baseline hazard is approximated using splines. The 'joineR' package [21], also within R software, fits an alternative parameterisation of the joint model, referred to as the 'random effects' (RE) parameterisation, where interest lies in modelling the association between the survival process and the latent random effects of the longitudinal process, representing deviations from the population average trajectory, as opposed to the true longitudinal response itself. Within this package, the survival process can only be represented utilising a Cox PH

model.

Detailed theory of longitudinal, survival, and joint modelling techniques is explored further within Chapter 2 of this thesis, where current limitations found within the literature, which this research alleviates, are highlighted throughout.

1.2.2 The Coxian Phase-type Distribution

The absorption times of any continuous time Markov process, with a single absorbing state, can be regarded as being phase-type distributed [22], where the distribution is defined by the transition intensity matrix of the Markov process which it represents. As such, the family of phase-type distributions is diverse; so much so that it can represent any positive distribution to an arbitrary degree of accuracy [13].

The practice of fitting phase-type distributions can be dated back to 1917, when Erlang [23], who was concerned with modelling the service times within a telephone exchange system, hypothesised that the total time spent within a queue could be considered to consist of a series of underlying states, through which individuals transitioned according to an exponential distribution before being served. This idea is often cited as the basis of modern queueing theory [24], and the practice of utilising a convolution of exponential distributions in this way, either in series or parallel, has been generalised to define a diverse family of distributions which are referred to as phase-type. The Coxian phase-type distribution is one such generalisation, which allows absorption to occur from any of the underlying transient states of the queue.

Phase-type distributions have previously been utilised to represent flow through various systems, such as patient flow through hospital [25,26] and students progression through university [27]. More recently, however, they have also been considered as a potential approach to model more typical survival analysis problems [28], where they can be employed to represent, for example, patients' 'flow' or progression through the states of a chronic disease [29]. Such applications, whilst novel, have motivated renewed activity within this research area. For instance, extensions to the phase-type methodology to allow for the incorporation of covariates within what is referred to as the phase-type regression model [30], as well as extensions for the inclusion of censored individuals [31], not typically encountered within standard queueing applications, have occurred relatively recently within the history of phase-type distributions.

Previous research has exhibited success at mapping the uncovered latent states from a fitted phase-type distribution onto distinct stages of the process which it represents. This is particularly true for the Coxian phase-type distribution, a popular

choice within the literature due to its underlying Markov process representing typical flow patterns. In doing so, inferences can be made from the estimated transition parameters on the rates of flow of individuals through the uncovered phases, providing further insight into the survival process, not ascertainable when utilising alternative distributions. This is advantageous when analysing survival data which is assumed to have an underlying Markov process, but where repeated measures were not collected on the individuals' transitions through the process.

At present, there exists only a small number of available approaches to fit phase-type distributions using standard statistical software, no doubt limiting their widespread application to survival analysis problems. A Bayesian approach to fit phase-type distributions is detailed within the R package 'PhaseType' [32], which is not capable of incorporating either covariate effects or censored individuals, reducing their applicability. EMpht [33] is another software approach which can be utilised to fit phase-type distributions which is capable of handling censored individuals but similarly cannot evaluate the effect of covariates on the system which the phase-type distribution represents. Indeed, there are currently no publicly available software packages which can handle both covariates and censored individuals.

Phase-type distributions are discussed in more depth within Chapter 3 of this thesis, where a number of methodological developments are also detailed to improve the suitability of phase-type distributions to represent typical survival processes.

1.3 Motivating Example

The research presented within this thesis is motivated, at least partially, by a study of individuals suffering from chronic kidney disease (CKD); a degenerative condition whereby an individual's kidney function gradually deteriorates over time. CKD often culminates in kidney failure, where haemodialysis (HD) treatment is necessary to fulfil the role of the ailing kidneys, before an eventual transplant is required.

Due to increasing incidence rates, where, for example, it has been estimated that 8.3% of individuals in England aged 16 or over will suffer from CKD by 2036, compared to 6.1% in 2011 [34], CKD is considered to be a prevailing challenge for healthcare providers [35]. Further, the disease has been observed to be particularly prominent amongst older individuals, where it is estimated that one in four women and one in five men aged between 64 and 75 are diagnosed with the disease [36,37], meaning incidents rates are increasing due to an ageing population [38].

Within this research, the novel joint modelling approach, which utilises the Coxian

phase-type regression model to represent the survival process, is used to analyse data collected from eight HD treatment centres across Northern Ireland. The aim of the analysis is to model the association between repeated measures collected on individuals' haemoglobin (Hb) levels, an emerging CKD biomarker, and their survival times. Within this analysis, the Coxian provides additional insight into the underlying stages of progression of end-stage renal patients, whilst also overcoming the aforementioned bias which can occur in survival parameter estimation by misspecifying the survival distribution.

1.4 Contributions

Throughout this research, a number of methodological developments are made, both to the area of phase-type distributions and to the area of joint modelling, significantly advancing each of the fields. The key advancements explored within this thesis are summarised below:

i A novel EM algorithm approach to fitting phase-type regression models is derived, with the aim of establishing a more stable method to fit such models and alleviate the identifiability issues which are well noted within the literature to impede current fitting procedures [39,40]. This approach, in comparison to standard Quasi-Newton (QN) and Nelder-Mead (NM) algorithm approaches, is advantageous as it also estimates what proportion of an individual's survival time is spent within each of the underlying states, allowing inferences to be made regarding quality of life; key information for both patients and clinicians.

ii The increased stability of this EM algorithm approach to fitting phase-type regression models is leveraged so as to relax the routinely imposed assumption that a covariate will have a constant effect across all transition intensities within the model. A new formulation of the standard Coxian regression model is presented, which instead allows (a) state-specific, (b) direction-specific and (c) transition-specific inferences to be made regarding the covariates' effects. In doing so, more information can be ascertained relating to how a covariate affects the system represented by the phase-type distribution.

iii The novel EM algorithm approach is also extended to allow for the inclusion of time-varying covariates, not considered within any phase-type model formulation within current literature. This advancement significantly extends the scope of phase-type distributions, as time-varying covariates are of increasing interest within many fields, particularly within medical statistics and epidemiology.

iv As detailed in Donnelly et al. [29], a two-stage approach to the joint analysis of longitudinal and survival data is preliminarily developed, utilising the Coxian phase-type regression model to represent the survival process. Within stage one of the model fitting procedure, a LME model is employed to overcome the measurement error which exists amongst the repeated measures of the longitudinal response, allowing an estimate of the 'true' response to be generated for each individual at their event time. Within stage two, these unbiased estimates are incorporated as predictors within the Coxian phase-type regression model, allowing the effect of the longitudinal process on the rates of transition through the underlying stages of the survival process to be quantified.

v Finally, a single joint likelihood approach to fit the new joint modelling specification, which utilises the Coxian phase-type regression model to represent the survival process, is derived for both the TLR and RE parameterisations of the joint model. Due to the flexibility of the Coxian phase-type distribution, in terms of its ability to represent any positive distribution, this new model overcomes previous limitations associated with fully parametric representations of the survival process, where bias can be introduced to the estimates of the survival parameters if the underlying distribution is misspecified. Further, the additional insight into the survival process provided by the Coxian, by way of the uncovered states, is not available from alternative survival representations and is a key advantage of the novel joint models presented within this work.

1.5 Thesis Outline

The remainder of this thesis is outlined as follows:

Chapter 2 presents a review of the literature within the areas of longitudinal, survival and joint modelling, as well as detailing the significant methodological developments which have been made within these fields of research. Throughout this review, limitations of standard joint models, which are subsequently targeted within this research, are highlighted.

Chapter 3 begins with a review of previous literature pertaining to phase-type distributions and phase-type regression models, where the limitations of the models which are addressed within this research are discussed. Subsequently, the development of the new EM algorithm approach to fitting phase-type regression models is presented in full. This new model is then extended to allow the effect of the covariates to vary across the transitions, increasing the information which

can be obtained regarding the survival process under investigation. Within this Chapter, the phase-type regression model is also extended to an inhomogeneous case, allowing time-varying covariates to be included within the model, necessary to incorporate the longitudinal response within the joint likelihood. Two simulation studies are presented throughout, validating the new methodological advancements which are introduced.

Chapter 4 details the development of the new joint model formulation which employes the Coxian phase-type regression model to represent the survival process. A preliminary two stage approach is first explored, before the single joint likelihood approach is detailed in full for both parameterisations. The chapter concludes with a simulation study, validating the new joint modelling framework and illustrating its advantages over the standard joint models currently available within published software.

Chapter 5 explores the application of the newly developed joint model which incorporates the Coxian phase-type distribution to represent the survival process to data collected from Northern Ireland renal patients suffering from chronic kidney disease.

Chapter 6 concludes the thesis by providing a summary of the main findings of this research, as well as presenting some potential points of further work.

Chapter 2

Joint Modelling of Longitudinal and Survival Data

2.1 Overview

The analysis of longitudinal and survival data, both independently and simultaneously within a joint modelling context, constitutes a widely discussed topic within statistical literature. This chapter reviews both the techniques for independent analysis of the two processes, alongside the more sophisticated joint likelihood approach for the simultaneous estimation of both processes. With regard to joint modelling techniques, some limitations of the standard procedures, subsequently addressed within this research, are discussed throughout.

2.2 Longitudinal Data Analysis

Over the last 40 years, significant developments have been made to the area of longitudinal data analysis, motivated partially by the more frequent collection of repeated measures data within an increasingly digitalised society. For example, the more regular use of computers to compile and store patient health records has meant that the collection of longitudinal data within the medical field happens almost inadvertently. Longitudinal studies, in comparison to cross-sectional studies, are favourable due to their increased statistical power; they have the capability to model change over time at both an individual and population level [41, 42], as well as the capacity to distinguish between ageing and cohort effects, which can otherwise become confounded [43], making longitudinal studies advantageous in terms of the type of statistical questions they can address.

On the other hand, the analysis of longitudinal data can prove challenging; such studies possess a number of unique features which make typical regression techniques redundant, and instead require specialist approaches in order to insure that valid inferences can be drawn [44]. Along with the improved capacity to collect and store repeated measures data, computational advancements have also aided the evolution of more sophisticated methodological techniques, necessary to draw such valid inferences from repeated measures data, meaning longitudinal studies have become increasingly popular within the fields of medicine, epidemiology and public health over the last 30 years [45].

2.2.1 Challenges Inherent to Longitudinal Data

Many of the problems faced when analysing longitudinal data stem from its multilevel hierarchical structure [46], which introduces naturally occurring heterogeneity

to the data meaning it can no longer be assumed independent and identically distributed. The repeated measures are instead clustered by individual as, intuitively, the observations collected on one individual are more alike to each other than to the observations collected on a second individual; this is due to the effect of unobserved, individual-specific characteristics which act upon the observed response.

Commonly, therefore, two sources of variation are considered when modelling longitudinal data: the between-individual variation, which represents the extent of the differences between individuals, and the within-individual variation, which represents how alike the observations made on a single individual are [46]. Figure 2.1 illustrates the multilevel nature of longitudinal data. As the independence assumption of ordinary linear regression is violated, techniques based upon the analysis of variance (ANOVA) paradigm [47] can potentially introduce bias, compromising their validity.

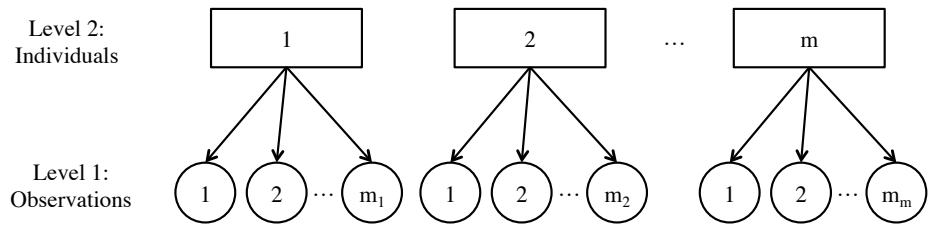


Figure 2.1: Diagram illustrating the multilevel nature of longitudinal data

Another frequent complication encountered within the analysis of longitudinal data is missingness. Missing data occurs when, for any reason, some individuals within a study do not have a complete set of observations. This can be due to either an individual skipping a scheduled observation appointment or, more commonly (and with a more significant impact), an individual dropping out of a study altogether before its completion. With regards to the former, 'skipped observation' missingness, it is instinctive that it is more straightforward to compare individuals' rates of change over time when all individuals have the same number of observations recorded at the same time points which, indeed, is a requirement for some of the less sophisticated historical approaches discussed in Section 2.2.2. When such a condition is not satisfied, bias can be introduced if the missingness problem is ignored, or individuals (and potentially informative data) can be omitted from the study if the problem is incorrectly addressed, possibly rendering the retained sample unrepresentative of the true population.

Missingness due to dropout can pose a more significant challenge to the analysis

of longitudinal data, particularly when the missingness is not at random [48]. For example, in a clinical trial to test the effect of a new drug intervention compared to a placebo on some repeatedly observed disease marker, if individuals who receive the drug are more likely to experience some negative reaction, causing them to drop out of the study, then those who received the treatment (and experience the negative side-effect) become under represented within the sample, introducing bias.

Missingness is closely related to the final prominent feature of longitudinal studies: unbalanced data. Whilst in theory it may seem beneficial to make the same number of observations, at the same time points, with equal time-gaps between observations, for all individuals within a study, this is difficult to enforce. In practice, it is inevitable that some observations shall be postponed, or missed entirely, for reasons beyond control, causing unbalanced and unstructured data. More drastically, if interested in conducting a retrospective analysis on data that was not collected as part of a trial with a strict structure, methods of analysis which require balanced data cannot be relied upon to produce unbiased parameter estimates.

2.2.2 Historical Methods of Analysis

Historically, a number of approaches have been considered to overcome the problems inherent to longitudinal data and, as shall be discussed, the path to developing the modern techniques used today has been a long one. Indeed, a review of the literature reveals a haphazard timeline within which different modelling frameworks were developed synchronously, each contributing to the advancement of the methodological framework towards the development of the linear mixed effects model (LME) model [1]; one of the most sophisticated approaches employed today to analyse longitudinal data.

2.2.2.1 Derived Variable Analysis

Derived variable analysis is the most simplistic approach which can be employed to handle the intercorrelation that exists amongst the repeated measures collected per individual within a longitudinal study. It involves compiling the information provided by the multiple observations into a single summary variable, meaning that more straightforward analytical approaches can be employed to the now uncorrelated summary variables, of which there is only one per individual [49]. That is to say, the approach is based upon modifying the data so as to suit the standard cross-sectional methods of analysis, rather than changing the method of analysis itself. Common derived variables include: (a) an 'average across time' variable, whereby the repeated measures are used to generate an average for each individual [50], (b) a 'change-score'

variable, where the change in response between the individuals' first and last observations is calculated, and (c) an 'average rate of change' variable, giving the average change between successive observations of each individuals' response.

There are a number of disadvantages associated with the derived variable approach. Firstly, summarising the repeated measures into a single observation can sacrifice the validity of the model by removing potentially informative data from the analysis. For example, an individual whose repeated measures show a negative trend over time could generate the same overall average as an individual whose repeated measures follow a positive trend; vastly different response profiles can generate the same derived variable. In such scenarios, whilst sufficient data was collected through the repeated measures to observe differences within the individuals which may be significant, the distinctions were lost in generating the derived variables, compromising the legitimacy of the results produced by the analysis. A similar loss of information occurs when calculating the 'change score' variable, as all but the first and last observations per individual are disregarded, meaning the analysis is blind to anything that occurs within this time interval.

Calculating a change score derived variable also presents additional challenges. The approach relies on the observed data having a balanced structure, and is vulnerable to missing observations and dropouts. In order to reasonably compare the change over time for each individual, it is necessary that all individuals are observed for the same length of time, otherwise those individuals who are observed for a shorter period will have less of an opportunity for the change to be observed. The two most commonly employed mechanisms to handle dropout are completer analysis, where only those individuals who were fully observed are retained within the analysis, or last observation carried forward (LOCF), where the last recorded observation, irregardless of the time at which it was made, is utilised to represent the individual's profile at study completion. Completer analysis can yield biased results as the individuals retained within the analysis may not represent the target population, and LOCF introduces bias as change scores observed over different time periods are being directly compared without considering that longer time periods allow for greater changes to occur. LOCF can also impact estimates of the means and standard deviations when employed to impute missing data.

The uncertainty within a derived variable is proportional to the number of repeated measures used to derive it. This means that, in cases where the data is unbalanced and individuals have different numbers of observations, there is a different uncertainty for each individuals' derived variable, thus violating the homoskedasticity assumption for standard ANOVA techniques [47].

The derived variable approach to longitudinal data analysis has a long history; in 1936, for example, Woodman et al. [51] conducted an investigation into the effect of a high protein diet on the growth rate of pigs. Within the experiment, three feeding treatments were tested by way of both group and individual feeding, with baseline and weekly weight measurements recorded for each pig. Within the analysis, however, only the first and last recorded weights were utilised to calculate a change score variable, disregarding all intermediate observations. In 1938, Wishart [52], noted that this omission of observations may lead to potential loss of information, and instead proposed modifying the method of analysis so as to incorporate the raw repeated measures data, rather than the derived variables.

2.2.2.2 Repeated Measures Analysis of Variance

Repeated measures analysis of variance (rANOVA) was one of the first methods of longitudinal data analysis to consider utilising the full set of the raw repeated measures collected in their unaltered form [41], and it does so by extending the simple ANOVA paradigm to include a single, individual-specific, random effect. The approach is comparable to that of the randomised block design, described by Fisher [53], commonly utilised within cross-sectional analysis when there exists some grouping (or "blocking") factor present within the experiment.

A block effect is a categorical variable which, in some way, groups together observations into similar classes which are inherent within the experimental units, rather than imposed by experimental design; the blocks are typically of no intrinsic interest but rather a nuisance variable which needs to be controlled for. That is to say, in order for valid inferences to be drawn when such clustering exists, the source of this between-block variation must be identified and isolated so as to prevent it from either confounding with the parameter estimates of the model or being misidentified as residual error. The classic example of a randomised block design, presented by Fisher [53], was concerned with designing an experiment to compare different types of fertiliser. Within the investigation, each of the fertilisers was tested across a number of fields, where each field was partitioned into various plots so as multiple fertilisers could be tested within each field. When conducting the analysis and evaluating the performance of the fertilisers, it was considered that, naturally, there would exist inherent variability in the performance of the fertilisers between the fields due to the underlying baseline fertility of the fields themselves, which additionally may contribute to the response of interest. To overcome this, a blocking variable for each field was introduced, representing the effect of the baseline fertility, preventing it from confounding with the effect of the fertilisers. Similarly, then, a blocking effect could be

considered within a longitudinal study to represent the underlying influence of any unobserved, individual-specific features or characteristics which also have an impact on the response of interest [54].

The rANOVA model is a simplistic form of what is now commonly referred to as a random intercept model, within which individuals' trajectories are allowed to deviate from the overall population average by some individual-specific random effect, denoted b_i, which is considered fixed over time. Consequently, the total variation within the rANOVA model which is not explained by the covariates is considered to come from two sources: (i) the variation caused by the individual characteristics, denoted b_i, where it is assumed $b_i \sim \mathrm{N}(0, \sigma_b^2)$ and (ii) the residual error, denoted ϵ_{ij}, where $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_e^2)$.

There were a number of limitations associated with the early rANOVA approach. Firstly, it imposed a strict compound symmetric covariance structure on the data whereby it is assumed that the variances and covariances remained fixed over time [55], whereas intuitively it would be expected that the variance would increase and covariance would decrease as the time gap between observations gets larger. This restriction was later alleviated by Greenhouse and Geisser [56] to allow for more general covariance structures. Secondly, the model, at least before later developments proposed by Henderson [57], required a balanced data structure with an equal number of observations recorded at fixed observation points for each individual, and was vulnerable to the effects of missing values. In fact, individuals with missing values were often omitted from the analysis when employing this approach, creating bias [58] as discussed previously. Finally, the approach does not have the flexibility to model scenarios where individuals have unique rates of change over time, along with unique intercept values, as is often observed in practice, limiting their applicability to real-world data.

Nevertheless, the rANOVA approach was popular due to the relative straightforwardness of the calculations involved and the reliability of the results when fitted to a balanced dataset which observed the strict assumptions made regarding the covariance structure of the data. The scarcity of such scenarios in practice, however, meant that alternative approaches, such as a modification of the multivariate analysis of variance approach, were also considered and developed.

2.2.2.3 Multivariate Analysis of Variance

In order to overcome some of the limitations faced when employing an rANOVA approach, longitudinal research moved in the direction of multivariate analysis of variance; a technique historically utilised within cross-sectional analysis to model multiple, possibly correlated, response variables of interest [59, 60]. Similarities were observed,

for example by Box [61], between this notion, and that of modelling multiple observations per individual on the same response. Namely, that both scenarios have correlations which must be contended with. This idea was generalised by Potthoff and Roy [62], in 1964, who proposed a multivariate analysis of variance growth curve model for the analysis of repeated measures data.

Whilst more computationally intensive, the multivariate approaches to growth curve analysis are widely regarded within the statistical literature to be superior to the univariate approach due to their increased flexibility [63–65]. For instance, they relax the strict compound symmetric assumption from before, instead imposing no restrictions on the covariances, allowing the model to more faithfully represent the longitudinal process. However, there are limitations associated with these approaches as well. Complete balanced data is required and individuals with missing observations are often omitted from the analysis, creating bias. Also, as the multiple measures are considered to be distinct rather than repeated observations of the same variable, time varying covariates cannot be intuitively handled within the analysis; the model does not explicitly consider how the distinct variables relate to each other, and so the temporal aspect is not modelled [66].

2.2.3 Linear Mixed Effects Models

The most significant contribution to the analysis of repeated measures data came in 1982 when Laird and Ware [1] proposed what they referred to as the random effects model for longitudinal data analysis, which forms the basis of modern linear mixed effects (LME) models. Within their research, they extended previous methodology, originally proposed by Harville to analyse multilevel data [67], to the specific case of longitudinal data where observations were considered to be clustered by individuals, as illustrated previously within Figure 2.1. LME models, now the most common approach taken to analyse longitudinal data, have been utilised, for instance, in cancer screening to describe longitudinal changes of cancer biomarkers [68], in AIDS clinical trials to estimate viral decay rate parameters [69], in depression treatment studies [70] and diabetes health management evaluation studies [71], to give just a few examples.

2.2.3.1 Background

As discussed previously, hierarchical data presents significant challenges within its analysis due to the presence of naturally occurring heterogeneity. Whilst specific focus is given here to the longitudinal scenario whereby observations are clustered within individuals, in practice hierarchical clustering can occur for many reasons,

for instance patients can be clustered within hospitals or students clustered within schools. A significant contribution to the analysis of such clustered data came in 1977, when Harville [67] presented a maximum likelihood estimation approach for a variance components model with both fixed and random effects, offering a number of advantages. Namely, the approach had no requirement for balanced data and allowed explicit modelling of the within-cluster and between-cluster effects.

The benefits of the approach can perhaps best be illustrated by looking at the specific case of students nested within schools; a common problem discussed within the literature. Over the years, many studies have been conducted to investigate factors which influence students' performance, where the hierarchical clustering of students within schools must be contended with. In 1976, before the development of Harville's approach to clustered data, Bennett [72] investigated the effect of different teaching styles on pupil performance, with particular attention paid to whether there exists an interaction between teaching style and pupils' personality traits. Within the study, Bennett used questionnaires to establish an idea of educators' teaching styles and used cluster analysis to group together teachers of similar style. Likewise, students were assessed and clustered based on their personality traits. Subsequently, analysis of covariance was used to determine if there existed significant differences between the performances of students exposed to different teaching styles, with the results indicating that a 'formal' teaching style lead to greater academic improvement, evaluated via aptitude tests conducted at the beginning and end of the school year.

The results, although widely referenced, were criticised by Gray and Sutterly [73] who re-evaluated the research and instead determined that no valid conclusions could be reached based on the published evidence. They criticised, specifically, the methods of establishing the clusters of similar teaching styles and highlighted that not enough was done to control for external influences. Both issues, whilst presenting real problems to the analysis of such multilevel data in 1976, can be handled in a relatively straightforward manner by using the variance components approach proposed by Harville. In fact, in 1986, Aitkin et al. [74] discussed the issues encountered within previous school effectiveness studies and advocated for the use of the newly developed variance component models for the analysis of studies within which observations are clustered, referencing the earlier work on variance components models (and their estimation) of Laird [75], Dempster, Laird and Rubin [76] and Dempster, Rubin and Tsutakawa [77].

As mentioned previously, Laird and Ware [1] were the first to extend this methodology to the case of longitudinal data, with later contributions coming from Laird [75], who discussed the computation of the variance components utilising the newly devel-

oped EM algorithm, and Laird, Lange and Stram [78] who, further investigating the maximum likelihood computation, solidified the approach as a fundamental longitudinal data analysis technique.

2.2.3.2 Basic Concepts

Broadly speaking, the LME approach to the analysis of longitudinal data involves the building of a model within which the response variable of interest is defined by both fixed effects and unobserved, individual-specific random effects. Within this context, the fixed effects are the observed covariates within the model which are presumed to have a constant (or 'fixed') effect on the rate-of-change of the response variable across all individuals. Much like the predictor variables within an ordinary linear model, they explain the population-level variation within the response variable of interest.

The random effects, in contrast, represent any intrinsic characteristics of the individuals which are unobserved within the data but which have an effect on the response variable. Because the random effects are individual-specific, it is reasonable to assume that their influence is exerted over all observations made on the same individual, thus causing these observations to vary systematically from the overall population average in a similar way. The random effects represent the source of the heterogeneity within longitudinal data; observations made on the same individual are more alike to each other than to observations made on another individual because the two subjects are influenced by a different set of random effects. The random effects can be thought of as random variables as they are assumed to follow some specified distribution; most typically the normal distribution.

Considering the fixed and random effects mathematically, the observed response for the i^{th} individual, y_i, observed at time t_{ij}, relates linearly to the fixed effects, ϕ_i, the random effects, ω_i, and the residual error, ϵ_i, as given by:

$$\begin{aligned} y_i(t_{ij}) &= \phi_i(t_{ij}) + \omega_i(t_{ij}) + \epsilon_i(t_{ij}) \\ &= y_i^*(t_{ij}) + \epsilon_i(t_{ij}) \end{aligned} \tag{2.1}$$

for $i = 1, \ldots, m$, where m is the total number of individuals observed, and $j = 1, \ldots, m_i$, where m_i is the number of observations made on the i^{th} individual. The true longitudinal response at time t_{ij}, denoted $y_i^*(t_{ij})$, is given by considering only the fixed and random effects, thus removing the residual error. The vector of residual errors for each individual is assumed normally distributed, $\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \mathbf{R}_i)$, and it is commonly

assumed that $\mathbf{R}_i = \sigma^2 \boldsymbol{I}_{m_i}$, where $\boldsymbol{I}_{m_i}$ is an $(m_i \times m_i)$ identity matrix [1, 67].

The fixed and random effects can be defined in a number of ways to specify different special cases of the LME model. For example, considering the fixed effects to be given by various explanatory variables with corresponding regression parameters $\boldsymbol{\beta}$, as in a standard linear model, i.e. $\phi_i(t_{ij}) = \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta}$, and allowing the individuals' intercepts to vary from the overall population average by some individual-specific constant b_i, i.e. $\omega_i(t_{ij}) = b_i$, produces the rANOVA model discussed previously in Section 2.2.2.2 and given below:

$$y_i(t_{ij}) = \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + b_i + \epsilon_i(t_{ij}) \tag{2.2}$$

where $\boldsymbol{x}_i(t_{ij})$ is a $(1 \times r)$ row vector of r fixed effects observed at time t_{ij}, and $\boldsymbol{\beta}$ is an $(r \times 1)$ vector of the corresponding fixed effects parameters.

This rANOVA formulation is the simplest form of LME model and is often referred to as a random intercept model, within which individuals can deviate from the population average by their intercept only, meaning that their trajectories (or rates-of-change over time) are defined only by the population-level parameters, i.e. they are assumed constant across the population.

Allowing individuals to be influenced by multiple random effects instead allows for more sophisticated models to be built whereby individuals can vary from the population average in terms of multiple covariates, not just their intercept. Setting $\omega_i(t_{ij}) = \boldsymbol{z}_i(t_{ij})\mathbf{b}_i$ produces the most generalised formulation of the LME model, given by:

$$y_i(t_{ij}) = \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + \boldsymbol{z}_i(t_{ij})\boldsymbol{b}_i + \boldsymbol{\epsilon}_i(t_{ij}) \tag{2.3}$$

where $\boldsymbol{z}_i(t_{ij})$ is a $(1 \times p)$ row vector of explanatory variables observed at time t_{ij} defining the $(p \times 1)$ vector of random effects, $\mathbf{b}_i$. Depending on how this design vector of the random effects is defined, various special cases of the LME model can be specified. For instance, allowing $\boldsymbol{z}_i(t_{ij}) = 1$ reduces the generalised LME model (2.3) to the random intercept model (2.2). Letting $\boldsymbol{z}_i(t_{ij}) = \boldsymbol{x}_i(t_{ij})$ gives a special case referred to as the random coefficients model [46, 79] in which each fixed effect has a corresponding random effect. Most commonly within the literature two random effects are considered, allowing individuals to deviate from the population average in terms of their intercept and their rate of change over time, where $\boldsymbol{z}_i(t_{ij})\mathbf{b}_i = b_{i0} + b_{i1}t_{ij}$.

In matrix notation, the generalised model (2.3) for individual i, introduced by Laird

and Ware [80], is given by:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, m \tag{2.4}$$

where

$\boldsymbol{y}_i$ is an $m_i \times 1$ vector of the m_i observed responses for individual i,

$\boldsymbol{X}_i$ is an $m_i \times r$ design matrix of the r observed covariates for individual i,

$$\text{i.e. } \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_i(t_{i1}) \\ \boldsymbol{x}_i(t_{i2}) \\ \vdots \\ \boldsymbol{x}_i(t_{im_i}) \end{pmatrix}$$

$\boldsymbol{\beta}$ is an $r \times 1$ vector of the unknown population (or fixed effects) parameters,

$\boldsymbol{Z}_i$ is an $m_i \times p$ design matrix of the p random effects for individual i,

$$\text{i.e. } \boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{z}_i(t_{i1}) \\ \boldsymbol{z}_i(t_{i2}) \\ \vdots \\ \boldsymbol{z}_i(t_{im_i}) \end{pmatrix}$$

$\boldsymbol{b}_i$ is a $p \times 1$ vector of the unknown, individual specific random effects,

$\boldsymbol{\epsilon}_i$ is an $m_i \times 1$ vector of the residual error terms, where it is assumed $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{m_i})$.

Similarly to the residual errors, the random effects are assumed multivariate normally distributed, $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a $p \times p$ positive-definite covariance matrix of the individuals random effects. The observed longitudinal responses can be considered conditionally independent given random effects, $\boldsymbol{y}_i|\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,\ \sigma^2\boldsymbol{I}_{m_i})$, meaning the likelihood function can be given given by:

$$L(\boldsymbol{\theta}_y, \boldsymbol{\theta}_b; \boldsymbol{y}_i) = \prod_{i=1}^{m} \int f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}_y) f(\boldsymbol{b}_i; \boldsymbol{\theta}_b) \mathrm{d}\boldsymbol{b}_i \tag{2.5}$$

where:

$$
\begin{aligned}
f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_y) &= \frac{1}{(2\pi\sigma^2)^{\frac{m_i}{2}}}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{b}_i)\right\}, \\
f(\boldsymbol{b}_i;\boldsymbol{\theta}_b) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{D}|^{\frac{1}{2}}}\exp\left\{-\frac{\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i}{2}\right\}.
\end{aligned}
\tag{2.6}
$$

which has a corresponding log likelihood given by:

$$
\begin{aligned}
\log L(\boldsymbol{\theta}_y,\boldsymbol{\theta}_b;\boldsymbol{y}_i) = &-\frac{m_i}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{b}_i) \\
&-\frac{p}{2}\log(2\pi)-\frac{1}{2}\log(|\mathbf{D}|)-\frac{1}{2}\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i.
\end{aligned}
\tag{2.7}
$$

Alternatively, the general LME model, given by Equation (2.4), can be expressed in its marginal form:

$$
\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{\epsilon}_i^* \tag{2.8}
$$

where $\boldsymbol{\epsilon}_i^*$ denotes the total between- and within-individual variation, i.e. $\boldsymbol{\epsilon}_i^* = \mathbf{Z_i b_i}+\boldsymbol{\epsilon}_i$, which is assumed to be normally distributed with an expected value $\mathbf{0}$ and covariance $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'+\sigma^2\boldsymbol{I}_{m_i}$, and allowing $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'+\sigma^2\boldsymbol{I}_{m_i}$ means $\boldsymbol{y}_i$ an be considered to be multivariate normally distributed, $\boldsymbol{y}_i \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta},\, \boldsymbol{V}_i)$.

The individual random effects, $\boldsymbol{b}_i$, are latent variables; although not explicitly observed within the data they are still of interest as they affect the individual level variation and so are still estimated. This is done using an extension to the Gauss-Markov theorem, developed by Harville [81] in 1976, where an estimate of $\boldsymbol{b}_i$ for each individual i is given by:

$$
\boldsymbol{b}_i = \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}) \tag{2.9}
$$

2.2.3.3 Linear Mixed Effects Models in Literature

Since its initial development in the early 1980s, the LME model has enjoyed a great deal of advancement, with contributions and extensions coming from many different areas of statistical interest. In 1984, for instance, Stiratelli et al. [82] generalised the LME model for serial observations with a binary response, utilising the EM algorithm to

estimate the unknown parameters. The authors compared their approach favourably to that of an earlier two step method, developed by Korn and Whittemor [83], highlighting the newer approach's ability to utilise all available data to give more reliable results, where as the earlier approach required individuals with low response rates to be excluded from the analysis, potentially causing bias.

In 1990, Lindstrom and Bates [84] further developed the work of Harville and Laird and Ware when they introduced non-linear mixed effects models, allowing more complicated longitudinal trajectories to be modelled. Gibbons and Hedeker [85], in 2000, described the application of the LME model to complex residual error structures whereby the assumption $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{m_i}$ does not hold. Instead they considered different types of auto-correlated errors, defined by a parameter of autocorrelation. The standard two level LME model was extended to three levels by Gibbons and Hedeker [86], motivated by the idea of multi-centre longitudinal clinical trial data where observations are nested within individuals and individuals within clinical centres. They also presented a three level example of individuals nested within classrooms within schools.

2.3 Survival Analysis

Survival analysis is the area of statistical research concerned with analysing the time from a well defined origin until an event of interest occurs [87]. Such techniques have a wide variety of applications, where focus primarily lies in quantifying the effect of various covariates on this time-to-event. For instance, they have been utilised extensively within clinical trials to identify treatment interventions which significantly improve survival time [88], within engineering to estimate the lifetime of electrical components [89] and in addiction studies to investigate factors which affect the time to relapse after a period of abstinence [90], to give just a few examples.

2.3.1 Features of Survival Data

There are a number of features associated with survival data which make typical methods of regression analysis redundant, and instead require specialist approaches to be applied in order for valid inferences to be drawn. Firstly, it is commonly observed that the event times are not normally distributed, but rather positively skewed, meaning that methods of analysis based on the normality assumptions cannot be relied upon. Secondly, it is typical that not all individuals within a survival study experience the event of interest while under observation, either as a result of premature dropout from the study, or due to the event not occurring during the pre-specified observation pe-

riod. Ignoring such individuals can result in model mis-specification and may result in biased parameter estimates, and so they must instead be incorporated within the analysis in some way, despite their missing response variable. This is done through censoring.

Consider that the true event time for individual i is given by τ_i^*, and allow C_i to represent the censoring time, i.e. the last time point at which an individual is observed before being lost to followup. If $C_i < \tau_i^*$ then individual i is censored and their true event time, τ_i^*, is never observed; all that is known about it is that it occurs at some time after C_i. The observed data within a survival investigation thus consists of each individuals final observation time, $\tau_i = \min(\tau_i^*, C_i)$, along with an indicator variable, δ_i, which takes the value 1 if the final observation time corresponds to the true event time and 0 if it alternatively corresponds to a censoring time.

The censoring mechanism can be classified as either informative or non-informative, depending upon whether there exists a relationship between the probability of being censored and the survival process under investigation.

Informative censoring occurs when an individual drops out of a study for some reason which relates to the survival process. For instance, in a survival investigation which aims to evaluate the effect of a treatment intervention on the time to death in cancer patients, informative censoring occurs if an individual prematurely drops out of the study due to a negative side effect of the treatment. When censoring is informative, the failure rates of those individuals observed within the study are significantly different from the failure rates of those who have dropped out [91].

Non-informative censoring occurs, in contrast, when an individual withdraws from a study for reasons unrelated to the survival process, for example due to relocation. The failure rates of such individuals are not considered to vary significantly from those individuals observed within the study

Much like when observations are missing not at random (MNAR) within a longitudinal study, as defined by Rubin [48], the options for dealing with informative right censoring are somewhat limited [19]; the observed data does not provide enough information to appropriately model that information which is missing. Consequently, the majority of literature focuses on cases where censoring is considered non-informative.

The final unique feature of survival data relates to conditioning; when estimating individuals' survival time, it has to be considered that the probability of survival changes as time progresses. For example, if the average estimated survival time of

individuals infected with some disease is seven years from the point of diagnosis, it has to be considered that an individual who is still alive at a time point six years after diagnosis is going to have an increased probability of surviving for seven years than an individual who is alive at only one year after diagnosis. Therefore, it is important to obtain survival probabilities that are conditional on the current status of the individual.

2.3.2 Survival Distributions

An observed event time within a survival study can be thought of as a continuous random variable drawn from some underlying distribution, defined by a probability density function. Two additional functions are often also employed within survival analysis to summarise the time-to-event data: the survivor and hazard functions [92], as described below.

Hazard Function, h(t)

The hazard function, denoted $h(t)$, represents the instantaneous risk of the event of interest occurring within the interval $[t,\ t+\delta t)$, where δt represents some infinitesimal variation in t, conditional upon the individual having survived until time t. The hazard function is given by:

$$h(t) = \lim_{\delta t \to 0} \left(\frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right) \tag{2.10}$$

Survivor Function, S(t)

The survivor function, denoted $S(t)$, represents the probability that the event of interest occurs after some specified time t, and is given by:

$$\begin{aligned} S(t) &= P(T \geq t) = 1 - F(t) \\ &= 1 - \int_0^t f(u)du \\ &= \exp\left\{ - \int_0^t h(u)du \right\} = \exp\Big\{ - H(t) \Big\} \end{aligned} \tag{2.11}$$

where $f(\cdot)$ denotes the probability density function, $F(\cdot)$ the cumulative distribution

function and $H(\cdot)$ is referred to as the cumulative hazard function.

The survivor function is an intuitive way to consider how censored individuals can be incorporated within the analysis; whilst their exact event time may be unobserved, it is at least known to occur at some point beyond their censoring time i.e. $\tau_i^* > C_i$.

Density Function, f(t)

The probability density function, which can be interpreted as the rate of change of the cumulative distribution function, is given by the product of the survivor and hazard functions, as shown:

$$\begin{aligned} f(t) &= \frac{P(t \leq T < t + \delta t)}{\delta t} \\ &= S(t)h(t). \end{aligned} \tag{2.12}$$

The expressions for the hazard, survivor and density functions depend upon the assumed distribution of the event times, where the most commonly employed distributions are the exponential, Weibull, gamma and log-normal due to their ability to represent positively skewed data. However, as outlined within Chapter 1, this research focuses instead on utilising the Coxian phase-type distribution to represent the survival process, where the Coxian is discussed fully within Chapter 3.

2.3.3 Fully Parametric Survival Models

When fitting a survival model under the assumption that the event times follow a specified underlying distribution, the model is said to be parametric; both the distribution of the event times and the effect of the covariates are described by parameters which are estimated during the model fitting procedure. When employing such a parametric approach, the likelihood function of the survival process is given by:

$$L(\boldsymbol{\theta}_\tau; \tau) = \prod_{i=1}^{m} \Big\{ S(\tau_i) \Big\} \Big\{ h(\tau_i) \Big\}^{\delta_i} \tag{2.13}$$

where the censoring process is assumed independent of the survival process and $S(\tau_i)$ and $h(\tau_i)$ are defined according to the assumed underlying distribution. Note that censored individuals only contribute to the likelihood through their survivor function, $S(\tau_i)$, as their event time is unobserved.

There are two common parametric approaches to the analysis of survival data; proportional hazards (PH) models and acceleration failure time (AFT) models, which are described briefly in Sections 2.3.3.1 and 2.3.3.2 below.

2.3.3.1 Proportional Hazards Models

Within statistical literature, PH models are perhaps the most common approach taken to analyse time-to-event data, where they quantify the effect of various covariates on the hazard of experiencing the event of interest. Consider, for example, a clinical trial designed to compare the effect of two treatment interventions on individuals' survival. If $h_1(t)$ represents the hazard function of those individuals who receive treatment 1, under the PH model formulation it is assumed that there exists some constant, denoted ξ, such that the hazard function of those individuals who receive treatment 2 is given by $h_2(t) = \xi h_1(t)$. As ξ is constant, this imposes a proportional hazards assumption upon the model; the hazard ratio for the two groups shall remain proportional over time:

$$\frac{h_2(t)}{h_1(t)} = \xi. \tag{2.14}$$

This constant ξ represents the ratio of the hazard of experiencing the event for individuals in treatment group 2, relative to the hazard of individuals in treatment group 1. The PH model can be generalised to quantify the effect of a vector of covariates, both binary and continuous, on the hazard of experiencing the event of interest by parameterising the hazard ratio such that $\xi_i = \exp\{\boldsymbol{w}_i'\boldsymbol{\gamma}\}$ for individual i, where $\boldsymbol{w}_i$ represents a vector of covariates with corresponding regression parameters $\boldsymbol{\gamma}$. In doing so the covariates are assumed to have a multiplicative effect on the hazard of death, where the hazard function for individual i is given by:

$$h_i(t) = h_0(t) \exp\left\{\boldsymbol{w}_i'\boldsymbol{\gamma}\right\} \tag{2.15}$$

where $h_0(t)$ represents the baseline hazard.

Despite their popularity within the literature, a disadvantage of PH models is that it is common for the proportionality assumption to be violated, meaning the inferences drawn from the models cannot be relied upon. AFT models, discussed in Section 2.3.3.2, are a commonly employed alternative when the PH assumption does not hold.

2.3.3.2 Acceleration Failure Time Models

Acceleration failure time (AFT) models are an alternative to PH models which can be employed when the PH assumption does not hold. Instead of evaluating the effect of the explanatory variables on the hazard of experiencing the event of interest, AFT models alternatively consider the effect of the explanatory variables directly on the survival times.

Consider, once again, a clinical trial designed to compare the effect of two treatment interventions. If $S_1(t)$ represents the survivor function of those individuals who receive treatment 1, under the AFT model formulation it is assumed that there exists some constant, again denoted ξ, such that the survivor function of those who receive treatment 2 is given by $S_2(t) = S_1(\xi t)$.

Within this context, ξ is referred to as the acceleration factor, where:

- $\xi > 1$ indicates an acceleration towards death (and thus a shorter survival time) in treatment group 2, relative to group 1,
- $\xi < 1$ indicates a deceleration towards death (and thus a longer survival time) in treatment group 2, relative to group 1.

As within the PH model, the AFT model can also be generalised to quantify the effect of various covariates, this time on the individuals' survival times. The generalised survivor function for the i^{th} individual under the AFT model formulation can be expressed by:

$$S_i(t) = S_0\left(t \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}\right) \tag{2.16}$$

where S_0 is the baseline survivor function and where ξ has this time been parameterised such that $\xi_i = \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}$, with $\boldsymbol{w}_i$ representing a vector of covariates and $\boldsymbol{\gamma}$ their corresponding AFT regression parameters.

2.3.4 Semi-parametric Survival Models

When fitting a fully parametric PH model, discussed within Section 2.3.3, the effect of the covariates, given by $\boldsymbol{\gamma}$, and the baseline hazard, $h_0(t)$, are estimated separately, with the covariate parameters being estimated first and subsequently utilised to give an estimate of $h_0(t)$ [87]. However, it can alternatively be considered that the PH

model, given by Equation 2.15, can be expressed as a linear model for the logarithm of the hazard ratio, as shown:

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \boldsymbol{w}_i'\boldsymbol{\gamma}. \tag{2.17}$$

As it is not necessary to estimate $h_0(t)$ in order to make inferences regarding the effect of the covariates on the relative hazard ratio, $h_i(t)/h_0(t)$, it is therefore not necessary to make any assumptions regarding the distribution of the survival times. The model is called semi-parametric as the parameters which define the underlying distribution of the survival times are not estimated during the model fitting procedure. Instead, the covariate parameters of this Cox PH model are estimated by maximising the partial likelihood, given by [2]:

$$L(\boldsymbol{\theta}_\tau;\tau) = \prod_{i=1}^{m}\left(\frac{\exp\left\{\boldsymbol{w}_i'\boldsymbol{\gamma}\right\}}{\sum\limits_{l\in R(\tau_i)}\exp\left\{\boldsymbol{w}_l'\boldsymbol{\gamma}\right\}}\right)^{\delta_i} \tag{2.18}$$

where $R(\tau_i)$ is the set of all individuals who have not experienced the event of interest at time τ_i, referred to as the risk set. While fully parametric models can be limited in terms of their scope due to the necessity of appropriately capturing the underlying distribution of the survival times, the semi-parametric Cox PH model can be employed to survival times according to any distribution, due to unrestricted functional form of the hazard.

2.3.5 Time-varying Survival Models

A well noted limitation of the previously discussed PH and AFT models is that they do not allow for time-varying covariates to be incorporated within their analysis, potentially limiting their scope. Indeed, with the increased collection of longitudinal data, as discussed within Section 2.2, it is becoming increasingly common for repeated measures to be made on various covariates relating to an individual's health condition throughout the observation period of a survival study. Omitting these observations from the survival model can result in the loss of valuable information.

The first approach to incorporate repeated measures within a survival model was proposed by Andersen and Gill [93], who extended the Cox PH model to allow the hazard to be defined by:

$$h_i(t) = h_0(t) \exp\left\{ \boldsymbol{w}_i'(t)\boldsymbol{\gamma} \right\} \tag{2.19}$$

where it can be observed that the vector of explanatory variables, $\boldsymbol{w}_i$, now depends on time, t. The partial likelihood of this model is given in a similar form to that of the Cox PH model by:

$$L(\boldsymbol{\theta}_\tau; \tau) = \prod_{i=1}^{m} \left(\frac{\exp\left\{ \boldsymbol{w}_i(\tau_i)'\boldsymbol{\gamma} \right\}}{\sum_{l \in R(\tau_i)} \exp\left\{ \boldsymbol{w}_l(\tau_i)'\boldsymbol{\gamma} \right\}} \right)^{\delta_i}. \tag{2.20}$$

A limitation of this approach is that it requires the time varying-covariates to be fully observable for each individual during the entire time for which they were under observation [94]. Consider, for example, that an individual q experiences the event of interest at time τ_q, and that at this time there are 2 individuals, r and s, also within the risk set $R(\tau_q)$. Consequently, the contribution to the likelihood of individual q is given by:

$$\left(\frac{\exp\left\{ \boldsymbol{w}_q(\tau_q)'\boldsymbol{\gamma} \right\}}{\exp\left\{ \boldsymbol{w}_q(\tau_q)'\boldsymbol{\gamma} \right\} + \exp\left\{ \boldsymbol{w}_r(\tau_q)'\boldsymbol{\gamma} \right\} + \exp\left\{ \boldsymbol{w}_s(\tau_q)'\boldsymbol{\gamma} \right\}} \right)^{\delta_q} \tag{2.21}$$

where $\boldsymbol{w}_r(\tau_q)$ and $\boldsymbol{w}_s(\tau_q)$ represent the vectors of covariate values for individuals r and s, observed at time τ_q, i.e. at the event time of individual q. This means, then, that in order to fit the time-dependent Cox model it is necessary to observe the covariate values for all individuals at each event time at which they are at risk. Whilst this may be attainable for exogenous covariates, it is not typically possible for endogenous covariates, such as individual biomarker readings. When employing the time-dependent Cox model to such endogenous covariates there is typically a large portion of missingness within the data. Often the LOCF approach is used to impute these missing values, introducing bias. In addition to this, the time-dependent Cox model does not take into consideration that the repeatedly observed covariate may be prone to measurement error, further exacerbating this bias [95].

2.4 Joint Modelling of Longitudinal and Survival Data

In healthcare modelling, it is typical for longitudinal and survival data to be collected simultaneously and, in such scenarios, it is often observed that there exists an asso-

ciation between the two processes [96]. Indeed, this interrelationship has been widely discussed within statistical literature, where previous research has shown that independently modelling the longitudinal and survival processes, utilising the methods discussed within Sections 2.2 and 2.3, can lead to biased results if such a relationship exists [9, 97]. Joint modelling techniques, consequently, are a relatively recent statistical development which aim to model both processes simultaneously, utilising all available data relating to both longitudinal progression and survival outcome, in order to make valid inferences about the two processes.

Whilst joint modelling approaches are necessary, most obviously, when interest lies specifically in evaluating the association between some longitudinal biomarker and a related survival outcome (i.e. the longitudinal and survival processes are of equal importance), they are also required when only one of the processes is the target of investigation, but where the other has an impact which needs to be taken into consideration and controlled for. That is to say, joint modelling approaches are necessary to make valid inferences when interest lies in modelling [98]:

i. the longitudinal response, when there exists informative dropout during the observation period [99], or

ii. the survival process, when interest lies in incorporating some time-varying endogenous covariate measured with error, or

iii. the extent of the latent association which is assumed to exist between both the longitudinal and survival processes.

It is easy to envisage the potentially large scope of joint modelling techniques within the fields of medicine and public health. Consider, as an illustrative example, the case of an individual admitted to hospital suffering from some disease. During their hospital stay, it is typical that repeated measures are made through time on various biomarkers which change in a way that reflects the underlying health condition of the patient. Intuitively, the changes in these markers may possess some predictive potential of a related future event outcome, either through their true empirical values or, importantly, through their rate of change over time. For example, at a certain point in time it may be that two individuals have approximately the same level of a biomarker covariate, indicating that they will have a similar survival experience. However, when looking at the patients' histories through their repeated measures, it may be that one individual's biomarker levels have been rapidly declining, where as the other individual's levels have remained fairly constant. Whilst the dynamic nature of the marker suggests the former individual's health condition is quickly evolving, with

the latter's remaining fairly constant, a cross-sectional view of the biomarker does not incorporate this information within the model. Joint models, on the other hand, can incorporate the informative patient history, improving the reliability of the results.

2.4.1 Two-stage Approach

Modern joint modelling techniques often reference back to research conducted by Tsiatis et al. [5, 100] and Self and Pawitan [6, 101], who were interested in overcoming the bias within survival models which occurrs when dealing with time-varying covariates, and De Gruttola and Tu [8, 102] who, conversely, wanted to overcome the bias introduced to longitudinal models as a result of informative dropout. Each of these three pioneering ideas looked specifically at modelling the association between a longitudinally observed endogenous biomarker in HIV patients, alongside a corresponding survival process. Whilst each of their approaches varied slightly, they all proposed a two-stage methodology where, in stage one, the process which was of secondary interest, but considered to introduce bias into the primary model, was fitted, and, in stage two, an unbiased feature of this process was incorporated within the primary model, yielding unbiased estimates of the association parameters.

Tsiatis et al. [5,100], interested in modelling the association between changing CD4 cell counts and the related survival times in HIV patients, acknowledged two common problems when observing endogenous covariates and evaluating their relationship with survival. Firstly, the covariates are observed intermittently and so a full covariate history, necessary when fitting a time-dependent Cox model, is typically not available for each individual, and, secondly, it is common for the covariates to be observed with some degree of measurement error. As an alternative to the time-dependent Cox, a two-stage approach was proposed whereby, in stage one, a growth curve random coefficients model was employed to generate personalised trajectories using empirical Bayes estimation approaches, as originally described by Laird and Ware [80]. Within this model, each individual's CD4 cell count was described by an individual-specific intercept and slope and was given by:

$$y_i^*(t_{ij}) = \theta_{i0} + \theta_{i1} t_{ij} \tag{2.22}$$

where θ_{i0} represents the intercept, or initial CD4 cell count, and θ_{i1} the slope, or rate of change of CD4 cell count, for individual i.

In stage two, estimates of the individuals' true covariate values for every time-point at which they are at risk are incorporated into a time-dependent Cox model so as to evaluate their effect on survival. Furthermore, the authors also considered the

possibility of incorporating other features of the individuals' repeated measures trajectories as potential predictors within the survival model. For example, they examined whether the individual-specific slopes, θ_{i1}, as well as the true CD4 cell counts, may have a strong association with survival outcome by fitting a survival model given by:

$$h_i(t) = h_0(t) \exp\Big\{ y_i^*(t)\alpha_1 + \theta_{i1}\alpha_2 \Big\} \tag{2.23}$$

where α_1 is the regression parameter corresponding to the individuals' true CD4 cell count at their death time and α_2 is the regression parameter associated only with the individuals' slope.

Whilst it was concluded that the individuals' unbiased CD4 cell counts were a significant predictor of survival, in this case it was found that the individual-specific slopes did not significantly contribute to the likelihood. However, this idea of incorporating features of the random effects within a survival model, alongside or in place of the true longitudinal response, forms the basis of the random effects parameterisation of modern joint likelihood approaches, discussed fully within Section 2.4.3.

Self and Pawitan [6, 101], interested in modelling the association between repeated observations on the ratio of individuals' T4 to T8 cell counts and the time until AIDS diagnosis in HIV patients, proposed a similar two-stage approach. This time, within stage one, a more traditional LME-type formulation to represent the longitudinal process was utilised, similar to Equation (2.4), where the model was fitted using the method of least squares by conditioning on the random effects. In stage two, they employed a survival model based on the multivariate counting processes of Anderson and Gill [93], modified so as to assume a linear relative risk form for the effect of the longitudinal process on survival, due to the fewer distributional assumptions regarding the random terms within the model, as noted by Prentice [95]. Other (possibly time-varying) covariates were also included within the survival model, where, for these covariates, the standard multiplicative form was assumed. The resulting survival model was of the form:

$$h_i(t) = h_0(t) \exp\Big\{ \boldsymbol{w}_i(t)\boldsymbol{\gamma} \Big\} \Big(1 + \boldsymbol{y}_i^*(t)\alpha \Big) \tag{2.24}$$

where $\boldsymbol{w}_i$ is a vector of additional covariates with regression parameters $\boldsymbol{\gamma}$.

Similar two-stage approaches were proposed by Dafni and Tsiatis [103], who incorporated empirical Bayes estimates of the time-dependent covariate at each event time within a Cox PH model, and Boycott and Taylor [104], who used population

smoothing methods employing a LME model with a stochastic process to generate more accurate estimates of the longitudinal response at each event time which were then incorporated within a time-dependent Cox model.

The primary drawback of these two-stage approaches is that they assume non-informative dropout during the observation of the longitudinal response. If this is not the case, as is often found in practice, then bias shall exist within the longitudinal model fitted in stage one, compromising the validity of the results.

De Gruttola and Tu [102], also in 1992, had the similar intention of removing the bias of one process on the other, but were instead primarily focused on modelling the rates of change of individuals' longitudinal response, where some observations were missing due to informative dropout. Specifically, they noted that the number of repeated measures made on the time-dependent covariates for each individual is related to the individual's survival time, whereby those individuals who survive longer will have a greater number of repeated measures recorded. Modelling longitudinal response in the presence of such informative censoring situations has been previously discussed by Wu and Carroll [105], who proposed a likelihood ratio test for informativeness and employed a probit model to derive coefficients corresponding to the informative censoring process which could then be incorporated within a mixed effects model.

De Gruttola and Tu proposed an extension of this methodology to consider a wide range of growth curve models where there exists a relationship between survival outcome and longitudinal response. Focusing specifically on the association between CD4 cell count and survival outcome in HIV patients, they estimated the contribution of each individual to the likelihood of the LME model by conditioning on their random effects, $\boldsymbol{b}_i$, assuming that the probability of being censored was unrelated to the individual's event time and that this conditional probability was independent of the longitudinal process. Fitting the model parametrically, assuming normally distributed random effects and that the joint distribution of CD4 cell counts and survival times is multivariate normal, allowed the EM algorithm to be employed to give estimates of parameters which describe the relationship between longitudinal response and survival.

Whilst these two-stage approaches can successfully correct for the measurement error which is commonly encountered within a repeatedly observed covariate, as well as overcome the disadvantages associated with the periodic nature of endogenous covariates' observation scheme, where the LME model can be utilised to predict the 'true' covariate at any time-point, they are still prone to producing biased estimates [96,106]. For example, when fitting a LME model in stage one, the effect of the survival process on the longitudinal process is not taken into consideration, and subsequently, the po-

tentially informative history of the individuals' repeated measures is not incorporated within the survival model. As such, joint likelihood approaches were explored within the literature.

2.4.2 Joint Likelihood Approach

In order to overcome the aforementioned sources of bias which are encountered when employing either the independent or two-stage approaches, methodological advancements have been developed to instead consider both the longitudinal and survival processes simultaneously through a single joint likelihood [3, 4]. This technique has been shown within the literature to yield better estimates of the longitudinal and survival parameters by more appropriately modelling the association between both processes [9, 97].

Depending upon the target process of interest, as well as the assumed underlying relationship between the two processes, this joint likelihood can be factorised in different ways to specify different formulations of the same model. Although, as noted by Hogan and Laird [107], the joint likelihood approach is global in the sense that all joint formulations are valid for each process and that factorising the likelihood with a particular target in mind does not affect the models overall likelihood, merely the interpretation of the parameters. The three common formulations, as described by McCrink et al. [108], are given below:

$$\textbf{Selection Models: } [\mathbf{Y}, \boldsymbol{\mathcal{T}}, \mathbf{b}] = [\boldsymbol{\mathcal{T}}|\mathbf{Y}]\ [\mathbf{Y}|\mathbf{b}]\ [\mathbf{b}] \tag{2.25}$$

$$\textbf{Pattern-mixture Models: } [\mathbf{Y}, \boldsymbol{\mathcal{T}}, \mathbf{b}] = [\mathbf{Y}|\boldsymbol{\mathcal{T}}]\ [\boldsymbol{\mathcal{T}}|\mathbf{b}]\ [\mathbf{b}] \tag{2.26}$$

$$\textbf{Shared Parameter Models: } [\mathbf{Y}, \boldsymbol{\mathcal{T}}, \mathbf{b}] = [\mathbf{Y}|\mathbf{b}]\ [\boldsymbol{\mathcal{T}}|\mathbf{b}]\ [\mathbf{b}] \tag{2.27}$$

where $\mathbf{Y}$ represents the longitudinal process, $\boldsymbol{\mathcal{T}}$ the survival process and $\mathbf{b}$ the latent random effects.

Selection models [109, 110], within which the survival process is conditional upon the longitudinal process, which itself is dependent upon the random effects, are most commonly employed in scenarios where the survival process is of primary interest, but considered to be influenced or associated with a longitudinal process which must also be taken into consideration.

Conversely, pattern-mixture models [111] assume the longitudinal process to be conditional upon the survival process, which is dependent upon the random effects,

making these models particularly applicable to scenarios whereby the longitudinal process is of primary interest. They have historically been explored within the literature to overcome missingness when analysing two continuous variables [112], and later to model the dropout mechanism in repeated measures data [113]. Hogan and Laird [107] and Sousa [114] provide a comprehensive overview of selection and pattern-mixture models.

This research focuses primarily upon the shared parameter joint likelihood formulation, where both the longitudinal and survival process are conditional upon the random effects, which represent the latent association between the two process. Noting the limitations of the two-stage approaches discussed within Section 2.4.1, it was Faucett and Thomas [3] who, in 1996, first considered estimating the parameters of both submodels simultaneously under the shared parameter formulation, where they employed the Markov chain Monte Carlo (MCMC) method of Gibbs sampling to estimate the joint posterior distribution of all the unknown parameters. The methodology was illustrated using simulated data, where the joint approach gave unbiased estimates of the parameters compared to the independent fitting of both models.

In 1997, Wulfsohn and Tsiatis [4] instead proposed a maximum likelihood approach to simultaneously estimate the longitudinal and survival parameters through a single joint likelihood utilising the EM algorithm. A LME submodel was utilised to model the longitudinal response, and a Cox PH submodel to represent the survival process, where only the true value of the longitudinal response, $y_i(t)^*$, was considered as a possible predictor of survival in what is referred to as the true longitudinal response (TLR) parameterisation of the joint likelihood, discussed fully in Section 2.4.4.

Henderson et al. [9], in 2000, generalised this approach of Wulfsohn and Tsiatis in a number of ways, firstly by incorporating baseline covariates as additional predictors within the survival model. Secondly, they considered other features of the longitudinal trajectory which may have a strong predictive potential of the survival outcome, such as the individuals' deviations from the population average, represented by the random effects, similarly to Self and Pawitan [6, 101], rather than exclusively considering the estimated true longitudinal response, $y_i(t)^*$.

Within their generalised modelling framework, they proposed that the longitudinal and survival processes be related through a subset of the individual's random effects, $\omega_i(t) = \{\omega_{i1}(t), \omega_{i2}(t)\}$, where $\omega_{i1}(t)$ influences the longitudinal response and $\omega_{i2}(t)$) the survival outcome. These two subsets of random effects have a latent association which is described by the cross correlation between $\omega_{i1}(t)$ and $\omega_{i2}(t)$ which constitutes a latent bivariate Gaussian process.

So, for example, allowing the longitudinal response to be represented by a LME model of the form:

$$y_i(t_{ij}) = \boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + \omega_{i1}(t_{ij}) + \epsilon_i(t_{ij}) \tag{2.28}$$

where $\boldsymbol{x}_i(t_{ij})\boldsymbol{\beta}$ represents the trajectory of individual i at time t_{ij}, as described by the population-level fixed effects, and $\omega_{i1}(t)$ represents the additional influence of individual-specific random effects. Typically, these random effects are specified by $\omega_{i1}(t_{ij}) = b_{i0} + b_{i1}t_{ij}$, where b_{i0} represents the individual-specific intercept and b_{i1} the individual-specific slope, as described in Section 2.2.3.

The hazard of individual i at time t, then, is given by:

$$h_i(t) = h_0(t) \exp\Big\{\boldsymbol{w}_i'\boldsymbol{\gamma} + \omega_{i2}(t)\alpha\Big\} \tag{2.29}$$

where $\omega_{i2}(t)$ could represent any feature of the longitudinal response and where α represents the corresponding parameter estimate(s).

In the previous case described by Wulfsohn and Tsiatis, the true longitudinal response at time t was incorporated within the survival model, i.e. $\omega_{i2}(t)\alpha = \big(\boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + b_{i0} + b_{i1}t_{ij}\big)\alpha$. Additionally, however, Henderson et al.'s approach allows for increased flexibility where $\omega_{i2}(t)$ can be alternatively specified. For example, multiple features of the random effects can be considered by allowing $\boldsymbol{\omega}_{i2}(t)\boldsymbol{\alpha} = \big(\boldsymbol{x}_i(t_{ij})\boldsymbol{\beta} + b_{i0} + b_{i1}t_{ij}\big)\alpha_1 + b_{i0}\alpha_2 + b_{i1}\alpha_3$, which evaluates the effect of (i) the true longitudinal response, (ii) the individual-specific deviation from the population-average intercept and (iii) the baseline individual-specific deviation from the population-average slope, on survival. Henderson et al. employed the EM algorithm approach proposed by Wulfsohn and Tsiatis to maximise the joint likelihood and estimate the unknown parameters of both submodels.

With Faucett and Thomas [3], Wulfsohn and Tsiatis [4] and Henderson et al. [9] providing this joint likelihood framework, early developments focused on extending the approach for alternative submodels. For instance, Wu [115], in 2002, motivated by the common application of non-linear mixed effects models to represent variations in viral load of HIV patients, and wishing to overcome the potential measurement error and informative dropout which can occur therewithin, proposed a joint likelihood approach within which the non-linear mixed effects model was used to represent the longitudinal process and a Monte Carlo EM approach was utilised to estimate the unknown parameters.

Alternatively, consideration has been given to modelling the longitudinal process utilising a LME model with a stochastic component. Henderson et al. [9], alongside their generalisation of Wulfsohn and Tsiatis's joint likelihood approach, also considered incorporating a stationary Gaussian process within the longitudinal submodel, where an EM algorithm approach was used to maximise the likelihood. Similarly, Salah et al. [116] incorporated an integrated Ornstein-Uhlenbeck process within the longitudinal submodel, utilising a Bayesian approach to estimate the unknown parameters due to the complexity of the model. Spline based approaches to representing the longitudinal process have also been explored within a joint model framework; Brown et al. [117] utilised non-parametric cubic B-splines to model the longitudinal process, where the joint likelihood was maximised using the MCMC algorithm. A similar approach was explored by Rizopoulos et al. [16] who instead estimated the parameters using a ML formulation.

With regard to the survival process, whilst it is convention within standard survival analysis to leave the baseline of a Cox PH model unspecified, this can lead to an underestimation of the standard errors within a joint modelling framework [11]. In order to overcome this, a fully parametric survival model can alternatively be utilised to represent the survival process, such as those described within Section 2.3.3. It is noted within the literature, however, that this can potentially limit the range of baseline hazards which can be accurately represented [12]. Rizopoulos [19] proposed that the step function approach of Whittemore and Killer [118] could instead be employed to generate non-parametric estimates of the baseline, or spline based approaches such as those proposed by Rosenberg [119] and Herndon and Harrell [120] could similarly be implemented.

Within this research, the survival process shall be represented by the Coxian phase-type distribution; a fully parametric approach, described in full within Chapter 4. It is well noted within the literature that phase-type distributions can be utilised to approximate any positive distribution arbitrarily closely [121], therefore making them more robust than the standard parametric survival approaches more commonly utilised.

Another common focus within the literature relates to the Gauss-Hermite quadrature approach to numerically approximate the (typically multi-dimensional) integral of the joint likelihood with respect to the random effects. Rizopoulos [19] noted that the computational difficulty of this numerical integration "constitutes the main reason why joint models have not yet found their rightful place in the toolbox of modern applied statisticians." In an attempt to improve the model fitting procedure, Rizopoulos et al. [16] and Lin et al. [122] have explored the use of Lapalace approximations, due

to their being more computationally efficient, particularly in cases with high dimensionality of the random effects. Rizopoulos [17] has also proposed a pseudo-adaptive Gauss-Hermite approach to improve the computation time of the standard Gauss-Hermite.

2.4.3 Random Effects Parameterisation

Under the random effects (RE) parameterisation of the joint likelihood, the association between the two processes is represented solely by the random effects, $\boldsymbol{b}_i$, which are present within both submodels. Joint models which observe this RE parameterisation can be fitted using the 'joineR' package within R Software [21], where the longitudinal process is represented by a LME model and the survival process by a Cox PH model, as shown below.

Longitudinal Submodel:

$$y_i(t) = \boldsymbol{x}_i(t)'\boldsymbol{\beta} + b_{i0} + b_{i1}t + \boldsymbol{\epsilon}_i(t) \tag{2.30}$$

Survival Submodel:

$$h_i(t) = h_0(t) \exp\left\{\boldsymbol{w}_i{}'\boldsymbol{\gamma} + b_{i0}\alpha_0 + b_{i1}t\alpha_1\right\} \tag{2.31}$$

Under this parameterisation, inferences can be made on the effect of a deviation from the population-average intercept or population-average slope on the survival process. For example, an individual whose intercept is one unit higher than the population average shall have a hazard which is $\exp\{\alpha_0\}$ times that of the hazard of the population average, so long as all other covariates remain constant.

The joint likelihood of the longitudinal and survival processes under the RE parameterisation, marginalised over the random effects, is given by:

$$L(\boldsymbol{\theta}; \boldsymbol{y}_i, \tau_i, \delta_i) = \prod_{i=1}^{m} \int f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}_y) f(\tau_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_\tau) f(\boldsymbol{b}_i; \boldsymbol{\theta}_b)\ d\boldsymbol{b}_i \tag{2.32}$$

where:

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}_y) = (2\pi\sigma^2)^{-\frac{m_i}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\right\}, \tag{2.33}$$

$$f(\tau_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_\tau, \boldsymbol{\beta}) = \left(h_0(\tau_i) \exp\left\{ \boldsymbol{w}_i{}'\boldsymbol{\gamma} + b_{i0}\alpha_0 + b_{i1}\tau_i\alpha_1 \right\} \right)^{\delta_i} \tag{2.34}$$

$$\times \exp\left\{ -\int_0^{\tau_i} h_0(s) \exp\left\{ \boldsymbol{w}_i{}'\boldsymbol{\gamma} + b_{i0}\alpha_0 + b_{i1}s\alpha_1 \right\} ds \right\} \tag{2.35}$$

$$f(\boldsymbol{b}_i; \boldsymbol{\theta}_b) = (2\pi)^{-\frac{q}{2}} |\boldsymbol{D}|^{-\frac{1}{2}} \exp\left\{ -\frac{\boldsymbol{b}_i' \boldsymbol{D}^{-1} \boldsymbol{b}_i}{2} \right\}. \tag{2.36}$$

where $\boldsymbol{D}$ is the variance-covariance matrix of the random effects.

The Gauss-Hermite quadrature, or the pseudo-adaptive quadrature, is commonly employed to numerically approximate the integral with respect to the random effects [17, 19]. Further, for the integral with respect to time within $f(\tau_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_\tau)$, an analytical solution cannot be obtained and instead it is necessary to approximate these integrals. Commonly, this is done using a 7-point or a 15-point Gauss-Kronrod rule [123].

2.4.4 True Longitudinal Response Parameterisation

Within the true longitudinal response (TLR) parameterisation, unbiased estimates of the true longitudinal process, denoted $y_i^*(t)$, are incorporated within the survival submodel so as to allow explicit inferences to be made regarding the effect of the longitudinal response directly on survival. Joint models which observe this TLR parameterisation can be fitted using the JM package within R Software [14], where typically the longitudinal process is represented by a LME model and the survival process by either a PH or AFT model, allowing for various underlying distributions to be assumed. The submodels of the two processes, when survival is represented by a Cox PH model, are shown below.

Longitudinal Submodel:

$$\begin{aligned} y_i(t) &= \boldsymbol{x}_i(t)\boldsymbol{\beta} + b_{i0} + b_{i1}t + \boldsymbol{\epsilon}_i(t) \\ &= y_i^*(t) + \boldsymbol{\epsilon}_i(t) \end{aligned} \tag{2.37}$$

Survival Submodel:

$$\begin{aligned} h_i(t) &= h_0(t) \exp\left\{ \boldsymbol{w}_i{}'\boldsymbol{\gamma} + y_i^*(t)\alpha \right\} \\ &= h_0(t) \exp\left\{ \boldsymbol{w}_i{}'\boldsymbol{\gamma} + \big(\boldsymbol{x}_i(t)\boldsymbol{\beta} + b_{i0} + b_{i1}t\big)\alpha \right\} \end{aligned} \tag{2.38}$$

Under this model formulation, $\exp\{\alpha\}$ represents the relative increase in the hazard

at time t as a result of a one unit increase in the longitudinal response at the same point in time.

The joint likelihood of the longitudinal and survival processes under the TLR parameterisation, marginalised over the random effects, is again given by Equation 2.32, where $f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_y)$ and $(\boldsymbol{b}_i;\boldsymbol{\theta}_b)$ are as defined previously within Section 2.4.3, but where:

$$f(\tau_i,\delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau,\boldsymbol{\beta}) = \left(h_0(\tau_i)\exp\left\{\boldsymbol{w}_i{}'\boldsymbol{\gamma} + \big(\boldsymbol{x}_i(\tau_i)'\boldsymbol{\beta} + b_{i0} + b_{i1}\tau_i\big)\alpha\right\}\right)^{\delta_i} \times \exp\left\{-\int_0^{\tau_i} h_0(s)\exp\left\{\boldsymbol{w}_i{}'\boldsymbol{\gamma} + \big(\boldsymbol{x}_i(t)'\boldsymbol{\beta} + b_{i0} + b_{i1}s\big)\alpha\right\}ds\right\}. \quad (2.39)$$

As can be observed, the longitudinal fixed effects parameter, $\boldsymbol{\beta}$, is now incorporated within the survival submodel, complicating the maximisation of the likelihood as a closed form expression of $\boldsymbol{\beta}$ cannot be obtained.

2.5 Summary

This chapter begins with a review of the independent methods of analysis for both longitudinal and survival data, before exploring the motivation behind the development of joint modelling techniques, whereby both process are estimated simultaneously. Alongside a review of the core developments within the theory of joint modelling, a number of limitations of standard joint models are identified as targets within this research. Specifically:

i. Employing the Cox PH model to represent the survival process within a joint modelling framework results in the underestimation of the standard errors due to the semi-parametric nature of the model [11], which has motivated the development of alternative representations of the survival process,

ii. Despite their intuitive interpretations, fully parametric representations of the survival process are not readily employed within a joint modelling framework due to limitations on the distributional shapes which they can represent [12], bolstering the popularity of piecewise constant baseline hazards approaches, as well as spline-based techniques, to model the survival process [19],

iii. When interest lies in making predictions on survival outcome from a joint model,

previous literature has established parametric models to be advantageous in comparison to spline based and piecewise constant approaches.

Within Chapter 3, the Coxian phase-type distributions, and the associated Coxian phase-type regression model, is explored as a potential alternative representation of the survival process. Whilst phase-type distributions offer a number of unique features, overcoming these aforementioned limitations, they also constitute a growing area of statistical research themselves, and their incorporation within a joint models serves to further their applicability within medical statistics, as they can currently not incorporate repeted measures.

Chapter 3

Exploring the Coxian Phase-type Distribution as an Alternative Survival Model

3.1 Overview

This chapter begins by introducing the standard methodology of phase-type distributions and exploring previous literature within this area of statistical research. Particular attention is paid to the Coxian phase-type distribution due to its suitability for representing typical flow patterns and its previously established applicability to typical survival analysis problems. Phase-type regression models, an extension to standard phase-type distributions which allow for covariates to be incorporated, are also introduced and an overview of their previous applications is presented.

Within Section 3.3, a new methodological approach to fitting phase-type regression models is detailed. This introduces an EM algorithm approach which is utilised due to its increased stability when fitting standard phase-type distributions, compared to previously applied fitting procedures such as the Nelder Mead (NM) and Quasi-Newton (QN) algorithms. Further methodological developments to this new EM algorithm approach are discussed within Section 3.4, detailing alternative formulations of the model which allow more in-depth inferences to be drawn regarding the covariates' effects on the system. Finally, within Section 3.5, the phase-type regression methodology is extended to allow for the inclusion of time-varying covariates, significantly increasing the scope of the models which, in current literature, cannot handle such covariates. Simulation studies are presented throughout, validating each of the new methodological advancements.

3.2 Phase-type Distributions

Phase-type distributions are a diverse family of distributions which describe the absorption times of a finite state Markov process in continuous time with a single absorbing state [22], formulated by a convolution of exponential distributions, either in series or parallel. As described by Neuts [124], phase-type distributions are a mathematically tractable way to approximate any positive distribution to an arbitrary degree of accuracy, which makes them a potentially attractive approach to represent survival data. In fact, such an advantage has been previously noted by Faddy [125], who remarked that phase-type distributions are advantageous in their ability to represent a greater variety of hazard shapes, particularly compared to the standard exponential and Weibull distributions which are limited to constant or monotone hazards, respectively.

Consequently, phase-type distributions are explored within this research as a potential alternative representation of the survival process within a joint model formu-

lation. Their fully parametric nature overcomes the previously discussed limitations experienced with semi-parametric survival models, while their ability to represent any positive distribution to an arbitrary degree of accuracy overcomes the noted constraints of alternative fully parametric approaches. Further, as shall be explored within this chapter, phase-type distributions have a number of unique advantages, compared to standard survival models, in terms of the inferences which can be made on the system that they represent, providing further insight into the process under investigation.

With that said, the application of phase-type distributions to typical survival analysis problems, in comparison to their more common usage within queueing theory and flow modelling, remains relatively novel, with little exploration beyond that of Aalen [28], who first proposed their applicability to biostatistics survival problems. As such, extensions to phase-type distributions to facilitate such implementations have occurred more recently within their history. For instance, they were adapted to handle censored individuals in 1996 [31], and the phase-type regression model, developed to incorporate covariate effects, was first proposed in 2012 [30].

As such, this chapter begins by reviewing standard phase-type distributions, before exploring advancements to the methodology so as to overcome some of the more inhibiting limitations of the models. Through this, phase-type distributions can be established as a more suitable approach to represent the survival process within a joint model, later explored fully within Chapter 4.

3.2.1 Background

The practice of fitting phase-type distributions began in 1917, when Erlang [23], interested in modelling the service times within a telephone exchange system, considered that the overall time spent within a queue could be represented by a series of n identical exponential distributions, defined by rate parameter λ. Conceptually, this is tantamount to assuming that the individuals within the queue progress through an underlying Markov chain of n transient states, before service, where they are then considered to enter the single absorbing state of the system, denoted state 0. This idea is represented diagrammatically within Figure 3.1

Erlang's so called 'method of stages' forms the basis of modern queueing theory [24], and the idea of utilising a convolution of exponential distributions, in series or otherwise, has subsequently been generalised in many ways to define a highly versatile class of probability distributions, referred to as phase-type [22]. For example, the hypo-exponential distribution is one such generalisation which relaxes the assumption that the series of exponential distributions are identical, and instead allows the transition

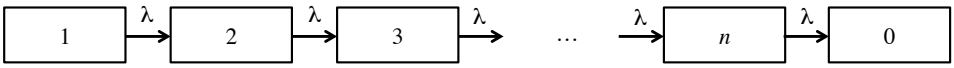


Figure 3.1: An illustrative representation of the Erlang distribution, where λ represents the rates of transition through the states of the underlying Markov process.

rates to vary across the underlying states, or 'phases' [126]. The Coxian phase-type distribution is a further generalisation which allows individuals to enter the absorbing state from any of the n transient states of the system, not just the final n^{th} state.

In general, phase-type distributions are therefore considered to represent the distribution of the absorption times of any continuous time, finite state Markov process with a single absorbing state, where the phase-type distribution is defined by the underlying Markov process which it represents. Consequently, there is a certain duality to the parameters of a phase-type distribution; while they can be considered holistically to represent the overall distributional shape of the absorption times, they can also be considered individually to represent the rates of flow through the underlying states of the Markov process which they represent.

Exploring these ideas mathematically, let us consider a continuous time Markov process, $\{S_t,\ t \geq 0\}$, defined on a state space $\{0, 1, ..., n\}$, where 0 is absorbing and $1, ..., n$ are transient. For each pair of states j and k, there exists an associated transition intensity, representing the instantaneous risk of transitioning from state j into state k, which is given by:

$$\begin{aligned} q_{jk} &= \lim_{\delta t \to 0} \Pr(S_{t+\delta t} = k | S_t = j)/\delta t \qquad j \neq k, \\ q_{jj} &= -\sum_{\substack{k=0 \\ k \neq j}}^{n} q_{jk}. \end{aligned} \tag{3.1}$$

where δt represents an infinitesimal time increment.

These transition intensities can be represented within a transition intensity matrix, $\mathbf{Q}$, which can be block partitioned to specify $\mathbf{T}$, a sub-generator matrix representing the instantaneous risk of transitioning among the transient states within the system, and $\mathbf{t}$, an exit vector representing the state-specific failure rates, or the rates of transitioning

into the absorbing state of the system. Thus, the overall transition intensity matrix is given by:

$$\mathbf{Q} = \left(\begin{array}{c|c} 0 & 0 \dots 0 \\ \hline \mathbf{t} & \mathbf{T} \end{array} \right). \tag{3.2}$$

where the zero terms along the top row indicate that state 0 is absorbing.

Additionally, a transition probability matrix, $\mathbf{P}(t)$, whose jk^{th} entry represents the probability of the process being in state k at time t, given that it was in state j at time 0, which, for a time homogeneous Markov process, satisfies the matrix differential:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q} \tag{3.3}$$

and can be calculated using the Kolmogorov differential equations [127] and is given by:

$$\mathbf{P}(t) = \exp\big\{\mathbf{Q}t\big\}. \tag{3.4}$$

The absorption times, τ_i, of such a Markov process with a single absorbing state are thus distributed according to a phase-type distribution, defined by two parameters: $\mathbf{p}$, a $(1 \times n)$ row vector indicating the probability of beginning the Markov process in each of the n transient states, and $\mathbf{T}$, the $(n \times n)$ sub-generator matrix of the system, as shown:

$$\tau_i \sim PH(\mathbf{p}, \mathbf{T}).$$

The probability density function of the phase-type distribution is given by:

$$f(t; \boldsymbol{\theta}_\tau) = \mathbf{p} \exp\big\{\mathbf{T}t\big\}\mathbf{t} \tag{3.5}$$

where $\mathbf{t}$ is the exit vector given by $\mathbf{t} = -\mathbf{T}\mathbf{1}$, $\mathbf{1}$ is an $(n \times 1)$ vector of ones and $\boldsymbol{\theta}_\tau$ represents the set of unknown parameters of the distribution; $\boldsymbol{\theta}_\tau = \{\mathbf{p}, \mathbf{T}\}$. Similarly, the survival probability of the phase-type distribution is given by:

$$S(t; \boldsymbol{\theta}_\tau) = \mathbf{p} \exp\big\{\mathbf{T}t\big\}\mathbf{1} \tag{3.6}$$

and the corresponding hazard by:

$$h(t;\boldsymbol{\theta}_\tau) = \frac{f(t;\boldsymbol{\theta}_\tau)}{S(t;\boldsymbol{\theta}_\tau)}. \tag{3.7}$$

Whilst there are clear similarities between phase-type distributions and Markov models, specifically in that they are both concerned with estimating the unknown transition parameters of the sub-generator matrix $\mathbf{T}$, they are different both in terms of the scenarios in which they can be utilised and in their primary target of inference.

When fitting a standard Markov model:

- the number of states within the system, as well as the criteria for determining to which state an individual belongs, is typically set out and defined in advance, based upon prior knowledge of the system under investigation,
- repeated measures are required, through time, to explicitly observe the current state of the system, where the likelihood is given by the product of the transition probabilities at all observation times over all individuals [128, 129],
- the objective of the model is to accurately represent the rates of flow amongst each of the states of the system; whilst the overall distribution of the absorption times can subsequently be represented by the estimated transition intensity matrix, the distribution is not considered during the parameter estimation process.

In comparison, when fitting a phase-type distribution:

- the number of states within the underlying Markov model is typically unknown and determined based upon that number which provides the best fit to the distribution of the absorption times, where likelihood ratio tests can be used to compare fits with increasing numbers of phases to identify the optimal number [130],
- no criteria is imposed upon the system to specify to which state an individual belongs; instead the fit, and number of underlying phases, is solely influenced by the data,
- only the single observation on an individual's absorption time is utilised to fit the model,
- the primary objective is to estimate the parameters which best represent the distributional shape of the absorption times, where making inferences from these parameters on the rates of flow of individuals through the underlying Markov process is a secondary additional benefit.

Phase-type distributions are therefore beneficial when employed to represent a process where there is an assumed but unobserved underlying Markov process, where interest may lie in making inferences regarding the rates of flow through this process in the absence of the repeated observations. For example, applying phase-type distributions to survival data can potentially uncover stages of a disease's progression and allow inferences to be made regarding how quickly individuals deteriorate through the stages of the disease. This uncovered information can provide insight into a patient's future quality of life, and can help inform treatment interventions; an individual who spends the majority of their survival time in the early stages of a disease before quickly transitioning through the more severe stages just before death will require different medical interventions than someone who will conversely spend the majority of their survival time in the more severe stages of the disease. Previously, Faddy and McClean [131], for example, have employed phase-type distributions to represent patient length of stay in hospital, utilising the uncovered states in this way to identify short, medium and long stay patients.

3.2.2 The Coxian Phase-type Distribution

The Coxian phase-type distribution [127], as previously mentioned, is a generalisation of the Erlang distribution which allows individuals to either (i) transition sequentially through the transient states of the underlying system with phase-specific intensities, denoted λ_j, or (ii) to transition from any transient state into the absorbing state, again with phase-specific hazards, denoted μ_j, as illustrated within Figure 3.2.

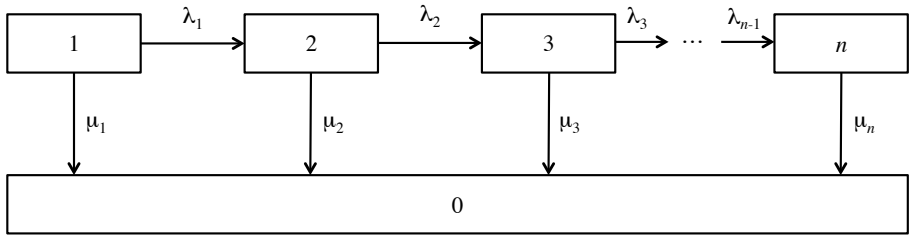


Figure 3.2: An illustrative representation of the Coxian phase-type distribution, where μ_j represents the rate of absorption from state j and λ_j represents the rate of transition from state j into state $j+1$.

The corresponding sub-generator matrix and exit vector of the Coxian phase-type distribution are given by:

$$\mathbf{T} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \dots & 0 \\ 0 & 0 & -(\lambda_3+\mu_3) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\mu_n \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_n \end{pmatrix} \tag{3.8}$$

where it is assumed that all individuals begin the process within the first state of the system, i.e. $\mathbf{p} = (1, 0, 0 \dots, 0)$.

The Coxian phase-type distribution is of particular interest within this research due to the fact that the typical flow patterns of a queue or survival process are well represented by the underlying Markov process of the Coxian. For example, when employing the Coxian to represent a failure process, the sequential transitions through the transient states can represent ageing, whilst the absorption transitions represent failure, with the uncovered phases representing different stages of the process [125]. When applied to the survival times of individuals suffering from some disease, it can be possible to map the uncovered states of the Markov model onto distinct stages of the disease's evolution, allowing meaningful inferences regarding the rates of flow through these stages to be obtained.

Previously, the Coxian phase-type distribution has been employed extensively to represent patient length of stay in hospital, where the transient phases of the system represent different stages of care, and patients can 'absorb' (i.e. leave the hospital) from any stage within the system [131–133]. They have also been employed, for instance, to represent operating and repair times of a device [134] and to fit heavy-tailed data within finance and insurance risk [135], to give just a few examples.

From a fitted Coxian phase-type distribution, Faddy [125], along with Marshall and McClean [133], previously considered estimating the proportion of failures which occur from each phase. The probability of experiencing the event of interest from each transient state, denoted ρ_j, is calculated by multiplying the probability of first surviving until the j^{th} state by the probability of absorbing from that state. A generalised expression for this probability is given by:

$$\rho_j = \left(\frac{\mu_j}{\mu_j + \lambda_j}\right) \prod_{h=1}^{j-1} \left(\frac{\lambda_h}{\mu_h + \lambda_h}\right), \quad \text{for } j = 1, ..., n \tag{3.9}$$

where $\lambda_n = 0$ and $\sum_{j=1}^{n} \rho_j = 1$.

The ordered event times can then be subdivided into the ratio $\rho_1 : \rho_2 : ... : \rho_n$, making it possible to determine n length of stay groups, W_j,:

$$W_j = \left\{ \tau^{(i)} : M \sum_{g=1}^{j-1} \rho_g < i \leq M \sum_{g=1}^{j} \rho_g \right\}, \tag{3.10}$$

where $\tau^{(1)}, ..., \tau^{(M)}$ are the ordered absorption times of the M individuals. This allows the identification of which individuals leave the system from which states, subdividing the data into groups with similar survival distributions. Further study of these groups can potentially provide more insight into what characteristics influence how individuals move through the system. This approach, however, imposes the assumption that all individuals who are absorbed from the first phase do so before any individuals are absorbed from the second phase, and so on. It does not allow for a scenario whereby one individual may quickly deteriorate through the system and absorb from the final phase faster than another individual may absorb from the first phase, for example. Conversely, within the EM algorithm approach to fitting phase-type distributions, discussed in detail later, the expected time each individual spends within each state is approximated within the E-step of the algorithm, overcoming this limitation.

Employing the forward Kolmogorov equation [136] to calculate the matrix exponential of the probability density function, given by Equation 3.5, whilst utilising the probabilities of absorption from each state, allows an analytic expression for the probability density function to be derived [137], as shown below, speeding up the fitting process:

$$f(t; \boldsymbol{\theta}_\tau) = \sum_{j=1}^{n} \rho_j \left(\sum_{k=1}^{j} C_{kj} (\lambda_k + \mu_k) e^{-(\lambda_k + \mu_k)t} \right) \tag{3.11}$$

where ρ_j is given by Equation 3.9 and:

$$C_{kj} = \prod_{\substack{h=1 \\ h \neq k}}^{j} \left(\frac{\lambda_h + \mu_h}{\lambda_h + \mu_h - (\lambda_k + \mu_k)} \right) \tag{3.12}$$

and where $\lambda_n = 0$.

3.2.2.1 Fitting Procedure

Phase-type distributions, in general, are well documented to suffer from identifiability issues; the same shape of distribution can be represented by multiple combinations of parameters, i.e. the distribution is non-singular [39], and the high number of unknown parameters which need to be estimated can often result in convergence to a local rather than global maximum [138]. Lang and Arthur [40], for instance, have previously investigated various approaches to evaluate the quality of the estimated parameters from both maximum likelihood and moment matching fitting procedures. Within their research, they noted the non-singular nature of phase-type distributions, showing that simulated distributions could be approximated by subsets of the phase-type distribution, and they concluded that a superior parameter estimation technique does not yet exist.

Coxian phase-type distributions, in comparison to more generalised phase-type distributions which allow transitions amongst all states within the underlying Markov process, alleviate this problem somewhat by reducing the number of parameters from $n^2 - n$ to $2n - 1$, whilst still retaining the ability to fit any positive distribution to an arbitrary degree of accuracy [33]. However, accurate methods for estimating their parameters is still considered an open problem [39, 40, 130] which, as noted by Marshall and Zenga [138], is caused by "the non-linear problem of fitting, the number of parameters to be simultaneously optimised and the non-unique representations of phase-type distributions". To date, various fitting procedures have been explored within the literature, with varying levels of success.

The Nelder-Mead (NM) algorithm is perhaps the most common approach, utilised by Faddy [130], Fackrell [139] and Marshall and McClean [140], for example, to maximise the likelihood of the model. Moment matching techniques have also been used [141], although Riska et al. [142] noted them to be ineffective at capturing the long tails of the distribution. Faddy [143] has also previously employed simple least squares along with a Quasi-Newton (QN) minimisation algorithm to estimate the unknown transition parameters, whereas Marshall and Zenga [138] have used the QN algorithm to instead perform maximum likelihood estimation. Marshall and Zenga [144], along with Payne et al. [145] have further discussed and compared different fitting procedures for the Coxian phase-type distribution, with Marshall and Zenga noting that whilst the QN algorithm is more effective computationally, the NM produces average parameter estimates which are closer to the simulated values. Asmussen et al. [33] instead

used the EM algorithm to estimate the unknown parameters within phase-type distributions, treating the single observations made on the individuals' absorption times as an incomplete observation on the full Markov process.

It has been extensively shown within phase-type literature that the maximisation of the likelihood is strongly influenced by the starting values of the unknown parameters; occasionally the problem converges towards a local maxima rather than the global maximum [146]. As the number of phases increases, so too does the number of unknown parameters and, consequently, the complexity of the maximisation problem, often increasing the number of initialisations which either fail to converge within the pre-specified criteria, or converge to a local maxima. Such convergence issues are symptoms of the previously discussed identifiability issues which are associated with fitting phase-type distributions and the extent of these failed convergences can vary depending on the algorithm employed to maximise the likelihood. Therefore, in order to determine the true best fit to the data, it is necessary to make multiple initialisations of the starting parameters, where the best fit is subsequently chosen based on a combination of the log-likelihood value, the shape of the estimated probability density compared to the empirical data and, where possible, clinical or professional input regarding expected rates of flow through the system under investigation.

3.2.3 The Coxian Phase-type Regression Model

Due, in part, to recent applications of the Coxian phase-type distribution within more conventional survival analysis settings, it has become of increased interest to incorporate covariates within the distribution so as to make inferences regarding their effects on the rates of flow through the underlying Markov process. However, a limited number of approaches to do so have been explored within the literature. Faddy [147], in 2009, considered the mean length of stay in the system to be dependent on various covariates using the log-link $\exp\left\{\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\right\}$, where $\boldsymbol{w}_i$ is a vector of covariates with corresponding regression parameters $\boldsymbol{\gamma}$. A fully Bayesian approach to this method was adopted by McGrory [148]. In 2012, Tang et al. [30] proposed the Coxian phase-type regression model, where instead the transition rate parameters were regressed on various covariate effects:

$$q_{jk} = q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\} \tag{3.13}$$

where q_{0jk} is the baseline intensity of transitioning from state j to k and $\boldsymbol{w}_i$ is a vector of covariates with corresponding regression parameters $\boldsymbol{\gamma}$. Theoretically, this same

approach can be taken to any phase-type distribution, not only one which is Coxian, where the probability density function is then given by:

$$f(t;\boldsymbol{\theta}_\tau) = \mathbf{p}\exp\left\{\mathbf{T}\exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}t\right\}\mathbf{t}\exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\} \tag{3.14}$$

where an AFT parameterisation is adopted, as shall be the case for the remainder of the novel research within this thesis.

This methodology, however, makes some restrictive constraints, specifically imposing the assumption that a covariate will have the same effect on each transition intensity. Consequently, within this model set-up, $\boldsymbol{\gamma}$ is interpreted as the acceleration effect of the covariate on the rate of flow through the entire system as a whole, where this effect is assumed to remain fixed over time. Whilst this is a standard assumption within conventional survival AFT models, it is in contrast to the standard approaches taken to estimating the effect of covariates within a Markov model, where transition-specific inferences regarding a covariate's effect can be made [149].

3.2.3.1 Comparison to Standard Survival Regression Models

As previously mentioned, one of the advantages of the Coxian phase-type distribution which is of particular interest within this research is its ability to represent any positive distribution to an arbitrary degree of accuracy [33,124]. In comparison, alternative survival distributions are noted within the literature to be restricted in terms of the distributional shapes which they can represent [125], limiting the scope of fully parametric survival models to cases where the data is known to observe a certain baseline distribution. The Coxian phase-type regression model, on the other hand, does not suffer from this same limitation, and instead can suitably represent any distributional shape by increasing the number of phases until the inclusion of an additional phase does not significantly improve the fit of the distribution. This idea can be best illustrated by way of an example.

Survival data was simulated according to an inverse-gamma distribution (chosen for its less-typical shape) with a single covariate effect. Subsequently, standard survival models which assume (i) an exponential, (ii) a Weibull, (iii) a log-normal and (iv) a log-logisic distribution were fitted to the data, with the estimated baseline distributions plotted alongside the simulated distribution within Figure 3.3. As can be observed, the Weibull and exponential distributions did not capture the true shape of the simulated distribution, whilst the log-normal and log-logistic failed to capture the distribution's peek.

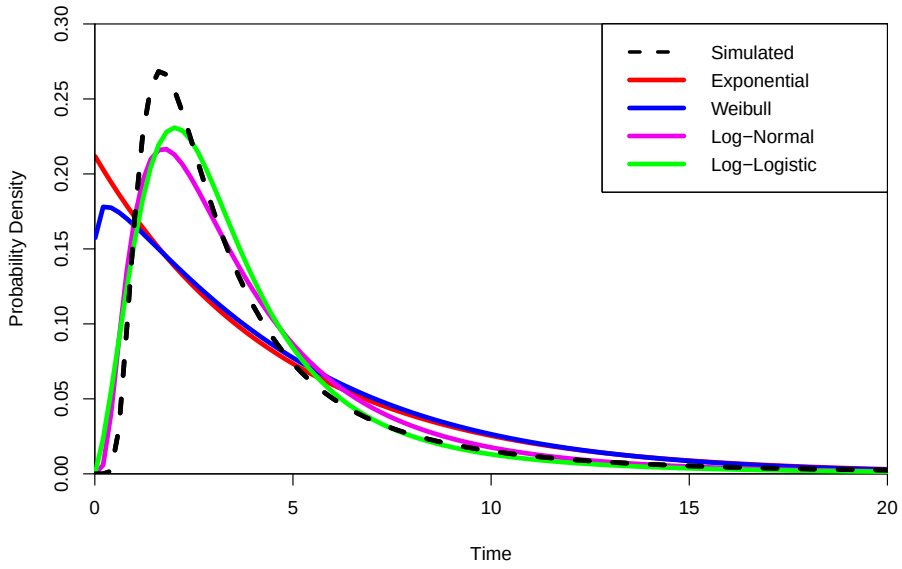


Figure 3.3: Estimated probability density functions from fitting (i) exponential, (ii) Weibull, (iii) log-normal and (iv) log-logistic survival models to the simulated data.

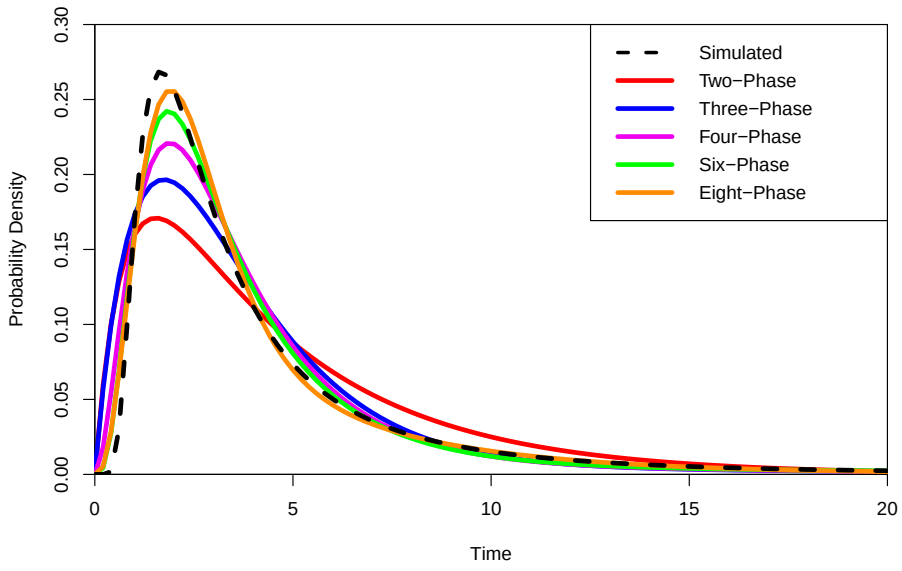


Figure 3.4: Estimated probability density functions from fitting the Coxian phase-type regression model with increasing numbers of phases to the simulated data.

Subsequently, the Coxian phase-type regression model was fitted to the same data, and by increasing the number of underlying phases it can be observed from the plotted density functions within Figure 3.4 that the fit of the Coxian distribution continued to improve until convergence towards the true simulated density. Thus, employing the Coxian phase-type regression model is beneficial within survival analysis as no prior knowledge of the baseline distribution of the data is necessary; its shape can always be captured by the Coxian.

3.2.3.2 Fitting Procedure

Beyond the research of Tang et al., there has been little published within the literature regarding the estimation of covariate effects within phase-type distributions; Faddy et al. [147] used built-in Matlab functions such as `fminsearch` to maximise the likelihood of their covariate dependent mean model. Tang et al. [30] instead extended a Bayesian approach, previously used by Ausin et al. [150], to fit the Coxian phase-type regression models.

To date, an EM algorithm approach to estimate the parameters of the phase-type regression model has not been investigated, despite the EM algorithm's increased stability compared to the NM and QN algorithms when fitting standard phase-type distributions. Consequently, within this research, such an EM algorithm approach is developed and, subsequently, the increased stability of the approach is leveraged so as to relax the restrictive assumptions made regarding the covariate effects. As such, this new EM algorithm approach significantly advances both the work of Asmussen et al. [33], which does not account for covariate effects, and previous phase-type regression models which do not allow transition-specific inferences of the covariate effects.

3.3 A New EM Algorithm Approach to Fitting Phase-type Regression Models

The EM algorithm, as formalised by Dempster et al. [76], is an iterative two step approach to obtain maximum likelihood estimates when handling data which is in some way incomplete. Whilst this missingness typically stems from incomplete observations in the traditional sense, the methodology is also applicable to cases where there exists some unobserved latent variable within the likelihood which must first be approximated [151], as is the case with the latent random effects within the LME model, for example. Similarly, when fitting phase-type distributions, the single observation on an individual's absorption time can be regarded as an incomplete observation on

the individual's full path through the underlying Markov process, where the missing data is comprised of the states the individual visits, and the time spent within these states [152].

The contribution to the complete likelihood of individual i is given by the product of the probability densities for all possible transitions through the underlying Markov model, as shown:

$$L(\boldsymbol{\theta}_\tau; \tau_i) = \prod_{i=1}^{m} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} f_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau) \tag{3.15}$$

where E_{ij} is the expected time individual i spends in state j, and $f_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau)$ is the probability density of individual i transitioning from state j into state k at this time. The parameters which maximise this likelihood also maximise the likelihood of the phase-type distribution which corresponds to the density function given by Equation 3.14. The density of $f_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau)$ can be expressed as the product of the hazard and survivor functions:

$$f_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau) = S_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau) h_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau)^{N_{ijk}} \tag{3.16}$$

where N_{ijk} represents the probability of transitioning from state j into state k.

The hazard of this transition is given by:

$$h_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau) = q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}\} \tag{3.17}$$

and the corresponding survivor function (where 'failure' in this case is exiting the state) is given by:

$$S_{ijk}(E_{ij}; \boldsymbol{\theta}_\tau) = \exp\left\{ - \int_0^{E_{ij}} q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}\} ds \right\} = \exp\Big\{ - q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}\} E_{ij} \Big\}. \tag{3.18}$$

where the vector of covariates, $\boldsymbol{w}_i$, is assumed time-invariant.

Equation 3.16 is comparable to Equation 2.13, where the density function of a typical fully parametric survival model is given by the product of the survivor function and hazard function taken to the power of the event indicator, δ_i. Therefore, when an individual is censored, $\delta_i = 0$ and only the survivor function contributes to the

density, whereas if the event is explicitly observed, $\delta_i = 1$ and both the survivor and hazard functions contribute to the density. On the other hand, when fitting a phase-type distribution using the EM algorithm approach, the individuals' paths through the underlying Markov process are unknown and it is not explicitly observed whether or not an individual makes each transition from state j to state k. This uncertainty is incorporated within the model via N_{ijk}, the probability of the transition from state j into state k occurring during the observation period.

For instance, if an individual is censored it is known that they do not enter the absorbing state during the observation period, i.e. $N_{ij0} = \delta_i = 0$ for all j. Conversely, if an individual is observed to experience the event of interest, it is known for certain that they absorb from one of the transient states in the system, i.e. $\sum_{j=1}^{n} N_{ij0} = \delta_i = 1$. Taking the survivor function to the power of N_{ijk} ensures the correct estimation of the density, where N_{ijk} are approximated within the E-step.

The contribution to the overall likelihood of individual i transitioning from state j into state k is thus given by:

$$\exp\Big\{ - q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\} E_{ij} \Big\} \Big(q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\} \Big)^{N_{ijk}} \tag{3.19}$$

and after incorporating the probability of beginning the process in each of the underlying states, given by $p_j^{B_{ij}}$, the overall likelihood of the complete Markov process for an n-phase type regression model can be expressed as:

$$L(\boldsymbol{\theta}_\tau; \tau_i) = \prod_{i=1}^{m} \left\{ \prod_{j=1}^{n} p_j^{B_{ij}} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} \exp\Big\{ -q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\} E_{ij} \Big\} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} \Big(q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\} \Big)^{N_{ijk}} \right\} \tag{3.20}$$

where:

q_{0jk} is the jk^{th} element of the baseline transition intensity matrix, $\mathbf{Q} = \left(\begin{array}{c|c} 0 & 0 \dots 0 \\ \hline \mathbf{t} & \mathbf{T} \end{array} \right)$, representing the transition intensities when $\boldsymbol{w}_i'\boldsymbol{\gamma} = 0$,

p_j is the j^{th} element of the phase-type initialisation vector, $\mathbf{p}$, representing the overall probability of beginning the Markov process in state j,

B_{ij} is the individual-specific probability that individual i begins the Markov process in state j,

E_{ij} is the total time individual i spends in state j,

N_{ijk} is the probability that individual i will make a transition from state j into state k at some point during the time for which they are in the system,

N_{ij0} is the probability that individual i belongs to state j at the moment before they experience the event of interest and transition into the absorbing phase,

$\boldsymbol{w}_i$ is a vector of covariate values for individual i, with corresponding regression parameters $\boldsymbol{\gamma}$, and,

m is the total number of observed individuals.

The approach taken to handle censored individuals extends upon that discussed by Olsson [31] for the incorporation of censored individuals within an EM algorithm approach to fitting standard phase-type distributions without covariate effects. It has been adapted within this research for the phase-type regression model.

The corresponding log-likelihood is given by:

$$\log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} B_{ij} \log(p_j) + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} -q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}\} E_{ij} + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} N_{ijk} \Big(\log(q_{0jk}) - \boldsymbol{w}_i' \boldsymbol{\gamma} \Big) \Bigg\}. \tag{3.21}$$

where the EM algorithm approach to maximising this likelihood is detailed below.

3.3.1 E-Step

When fitting phase-type distributions, the observed event times, i.e. $\tau_i = min(\tau_i^*, C_i)$, along with the event indicator, δ_i, and covariate values, $\boldsymbol{w}_i$, comprise the observed information for each individual. This means that the 'missing' data to be approximated for each individual within the E-step of the algorithm consists of:

i. the probability of beginning the process in each of the underlying states, B_{ij},

ii. the total time spent within each underlying state, E_{ij},

iii. the probability of a transition occuring between each pair of states, N_{ijk}, and

iv. the probability of absorbing from each of the underlying states, N_{ij0}.

Consequently, on any given iteration of the EM algorithm, the expected values of these latent variables are calculated, based on the current best estimates of the unknown parameters, $\boldsymbol{\theta}_\tau = \{\mathbf{p},\ \mathbf{T},\ \boldsymbol{\gamma}\}$. It should be noted that, for some specific phase-type distributions, such as the Coxian, it is assumed that all individuals start the process within the first state of the underlying system, meaning that both B_{ij} and $\mathbf{p}$ are known in advance, and are therefore not required to be predicted or estimated. For completeness, and to illustrate the methodology for the general case, the prediction and estimation of B_{ij} and $\mathbf{p}$, applicable in cases where this assumption is not made, is discussed here, however they shall not be required within subsequent applications which utilise the Coxian phase-type distribution.

The expected values of B_{ij}, E_{ij} and N_{ijk} on any given iteration of the EM algorithm are given by:

$$\mathbf{E}\Big[B_{ij} \mid \tau_i\Big] = \frac{p_j d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.22}$$

$$\mathbf{E}\Big[E_{ij} \mid \tau_i\Big] = \frac{c_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.23}$$

$$\mathbf{E}\Big[N_{ijk} \mid \tau_i\Big] = \frac{c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau) q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.24}$$

and, for individuals who experience the event of interest, the expected value of N_{ij0} is given by:

$$\mathbf{E}\Big[N_{ij0} \mid \tau_i\Big] = \frac{a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) q_{0j0} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.25}$$

where $N_{ij0} = 0$ for censored individuals. The values of $a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$ and $c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, as well as the elements of the vector $\mathbf{d_i}(\tau_i \mid \boldsymbol{\theta}_\tau)$, denoted $d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, are given by:

$$a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}\tau_i\Big\}\mathbf{e_j} \tag{3.26}$$

$$d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{e_j}' \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}\Big)^{\delta_i} \tag{3.27}$$

$$c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}u\Big\}\mathbf{e_j}$$

$$\times \mathbf{e_k}' \exp\left\{\mathbf{T}\exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}(\tau_i - u)\right\}\left(\mathbf{t}\exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}\right)^{\delta_i} du. \tag{3.28}$$

where $\mathbf{e_j}$ is the j^{th} unit vector and all power operations are performed in an element-wise fashion; i.e. $\left(\mathbf{t}\exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}\right)^0 = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones. The derivation of these expected values are presented within Appendix A. The Runge-Kutta numerical approximation can be utilised to predict $a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$ and $c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, as implemented by Asmussen [152] for phase-type distributions without covariate effects, or $a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$ and $d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$ can be calculated analytically using the current best estimates of $\boldsymbol{\theta}_\tau$, and $c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$ can be numerically approximated using the Gauss-Kronrod rule [123].

3.3.2 M-Step

Within the M-step of the algorithm, the log-likelihood is maximised and closed form estimates of the transition parameters are given by:

$$\begin{aligned}
\widehat{q}_{0jk} &= \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m} E_{ij}\exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}\}}, \quad \text{for } j = 1,\ldots,n\ ,\ k = 0,\ldots,n \text{ and } j \neq k \\
\widehat{q}_{0jj} &= -\sum_{\substack{k=0 \\ k\neq j}}^{n} \widehat{q}_{0jk}
\end{aligned} \tag{3.29}$$

and of the initialisation vector by:

$$\widehat{p}_j = \frac{\sum_{i=1}^{m} B_{ij}}{n}, \quad \text{for } j = 1,\ldots,n. \tag{3.30}$$

As closed form estimates of the covariate parameters can not be obtained, a one-step Newton Raphson process is used on the $(l+1)^{th}$ iteration to obtain updated estimates of $\boldsymbol{\gamma}$:

$$\widehat{\boldsymbol{\gamma}}^{(l+1)} = \widehat{\boldsymbol{\gamma}}^{(l)} - H\big(\widehat{\boldsymbol{\gamma}}^{(l)}\big)^{-1} S\big(\widehat{\boldsymbol{\gamma}}^{(l)}\big) \tag{3.31}$$

where the score vector, $S(\boldsymbol{\gamma})$, and Hessian matrix, $H(\boldsymbol{\gamma})$, are given by:

$$S(\boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} q_{0jk} \boldsymbol{w}_i \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}\} E_{ij} - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} \boldsymbol{w}_i N_{ijk} \right\}$$

$$H(\boldsymbol{\gamma}) = \frac{\partial^2}{\partial \boldsymbol{\gamma}^2} \log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} -q_{0jk} \boldsymbol{w}_i \boldsymbol{w}_i' \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}\} E_{ij} \right\}. \tag{3.32}$$

3.3.3 Simulation Study One

A simulation study was conducted to validate the new EM algorithm approach to fitting phase-type regression models and to compare the new methodology to previously employed NM and QN algorithm approaches. Within the study, to illustrate each of the three algorithm's robustness, or lack thereof, to an increasing number of phases and covariates, four scenarios were considered, where datasets were generated consisting of:

i. two underling phases, defined by baseline transition parameters $q_{010} = 0.05$, $q_{020} = 0.10$ and $q_{012} = 0.30$, influenced by a single continuous covariate generated from a uniform distribution, $w_i \sim \text{unif}(-3, 3)$, with corresponding regression parameter $\gamma = 0.2$,

ii. two underlying phases, defined by baseline transition parameters $q_{010} = 0.05$, $q_{020} = 0.10$ and $q_{012} = 0.30$, influenced by two covariates, one continuous: $w_{i1} \sim \text{unif}(-3, 3)$, and one binary: $w_{i2} \sim \text{unif}\{0, 1\}$, with corresponding regression parameters $\gamma_1 = 0.3$ and $\gamma_2 = -0.4$,

iii. three underlying phases, defined by baseline transition parameters $q_{010} = 0.02$, $q_{020} = 0.10$, $q_{030} = 0.20$, $q_{012} = 0.15$ and $q_{023} = 0.25$, influenced by a single continuous covariate, $w_{i1} \sim \text{unif}(-3, 3)$, with corresponding regression parameter $\gamma = 0.6$, and,

iv. three underlying phases, defined by baseline transition parameters $q_{010} = 0.02$, $q_{020} = 0.10$, $q_{030} = 0.20$, $q_{012} = 0.15$ and $q_{023} = 0.25$, influenced by two covariates, one continuous: $w_{i1} \sim \text{unif}(-3, 3)$, and one binary: $w_{i2} \sim \text{unif}\{0, 1\}$, with corresponding regression parameters $\gamma_1 = 0.6$ and $\gamma_2 = -0.2$.

The parameters for this simulation study were carefully selected to generate data which observes the shape of a typical survival distribution; i.e. positively skewed with a

long tail, in line with previous simulation studies conducted on the Coxian phase-type distribution [145].

For each scenario, 100 datasets were simulated using the 'actuar' package within R software [153], each consisting of 400 observations with approximately 20% censoring. A single initialisation was performed for each dataset using the NM, QN and EM algorithms, where convergence was considered to be achieved when the difference in the log-likelihood between two successive iterations was less than 1×10^{-4} and a maximum of 1000 iterations were permitted.

Within previous literature, a number of performance measures have been employed to evaluate the success of different algorithms at fitting phase-type distributions [138, 154], two of which are of interest here:

a. Rate of convergence (ROC)

   This is a measure of the number of initialisations which satisfy the convergence criteria within the maximum number of iterations, given by:

$$\text{ROC} = \frac{\text{number of convergences}}{\text{number of initialisations}} \times 100. \tag{3.33}$$

b. Rate of algorithm's success (RAS)

   This is a measure of the number of convergences which produce acceptable results, given by:

$$\text{RAS} = \frac{\text{number of acceptable results}}{\text{number of convergences}} \times 100 \tag{3.34}$$

   where an "acceptable result" is one for which the estimated parameters fall within an acceptable range (here considered to be between 0 and 10), as defined by Marshall and Zenga [138]. That is to say, not all convergences will suitably estimate the parameters of the distribution, where some may have converged at a local maxima rather than the global maximum. Additionally, certain approaches may be more prone to converging to a subset of the true phase-type distribution, as discussed by Lang and Authur [155]. The RAS, therefore, gives an indication of what proportion of those initialisations that converged resulted in a set of parameters which appropriately represents the true distribution of the data.

For further insight and ease of interpretation, these performance measures are combined within this research to introduce a third metric, the rate of successful conver-

gence (ROSC), which is a measure of how many acceptable fits are produced, relative to the total number of initialisations, which is given by:

$$\text{ROSC} = \frac{\text{number of acceptable fits}}{\text{number of initialisations}} \times 100 \tag{3.35}$$

This gives the more intuitive interpretation of the percentage of initialisations, rather than convergences, which result in an acceptable fit.

3.3.3.1 Results

The ROC, RAS and ROSC scores from the NM, QN and EM algorithm fits for the four simulated scenarios are given in Table 3.1.

The ROC scores indicate that the EM algorithm successfully satisfies the convergence criteria more often than either the NM or the QN algorithms for each of the four scenarios, suggesting it to be the most stable method of fitting Coxian phase-type distributions. In terms of convergence, the NM is both the poorest performing algorithm and the algorithm which is most affected by increasing number of phases and/or covariates. For instance, for the three-phase two covariate scenario, only 21% of the NM initialisations resulted in convergence, compared to 69% for the two-phase one covariate scenario; the addition of three extra parameters, resulted in more than a 66% decrease in the number of convergences. The QN algorithm, in contrast to the NM, seems largely unaffected by the inclusion of additional phases, whilst showing a slight decrease in ROC with additional covariates. The EM algorithm, conversely, appears more robust to the inclusion of additional covariates, but exhibited a slight decrease in ROC with additional phases.

From the RAS scores it can be observed that, for the two phase simulations, both the EM and QN algorithms have scores of 100%, indicating that all of those initialisations which resulted in convergence provided acceptable parameter estimates. The NM algorithm, on the other hand, has lower RAS scores, indicating that this approach is more susceptible to false convergence. Whilst the EM algorithm is more robust to increasing phases, maintaining its score of 100% for the three phase simulations, the QN begins to also suffer false convergence, although not to the same extent as the NM.

The ROSC score combines the ROC and RAS scores to reveal what percentage of the initialisations resulted in acceptable results, making it easier to infer which algorithm performs best overall. So, for example, whilst the QN algorithm may be least affected by increasing the number of phases in terms of ROC, considering this

along with the RAS scores, which suggest that the QN begins to experience problems with the quality of its convergences, provides a better insight. It can be observed that the EM algorithm overall performs significantly better than its competitors, with ROSC scores which are at least 20% higher than the next best algorithm for each of the simulated scenarios.

To illustrate how well each of the approaches uncover the true distribution of the absorption times, the mean probability density of the acceptable fits for the three phase simulations with one and two covariates are plotted in Figures 3.5 and 3.6 respectively.

For the three phase simulation with one covariate, it can be observed that the mean of all three algorithms suitably captured the shape of the distribution, although this was only the case for the NM and QN when the false convergences had been identified and removed. For the three phase simulation with two covariates, the QN and EM both captured the shape, with the EM performing marginally better with slightly smaller confidence intervals. On the other hand, those fits of the NM algorithm which were retained as "acceptable" did not suitably represent the shape; suggesting this approach requires more initialisations to yield results which accurately represent the distribution.

The mean estimates of the covariate parameters for the acceptable results are given within Table 3.2. It can be observed that the acceptable results of all three approaches were close to the simulated values, with the EM algorithm having the smallest standard errors. The EM algorithm is also superior in the sense that further investigation into those initialisations which converged is not required as they were all deemed acceptable. In comparison, even after further intervention to remove the false convergences, the QN and NM still had larger standard errors compared to the EM algorithm.

Whilst not a primary focus of this simulation study, it can be noted that the improved accuracy of the EM algorithm comes at the expense of longer computational times. Previously Payne et al. [145] noted that the timings for a single convergence of the Coxian phase-type distribution varied significantly depending on the software employed to fit the models, with MATLAB functions taking approximately 529 seconds (8.8 minutes) to reach convergence for a three-phase Coxian. Whilst the EM algorithm approach takes longer for a single initialisation, it has the advantage of not needing as many initialisations to produce an 'acceptable result'.

Overall, the newly developed EM algorithm approach is shown to improve upon the NM and QN algorithms within each performance assessment, identifying it as the better approach.

Table 3.1: Summary of the ROC, RAS and ROSC values over 100 simulations for the NM, QN and EM algorithms.

No. of Phases	No. of Covs	ROC			RAS			ROSC		
		NM	QN	EM	NM	QN	EM	NM	QN	EM
2	1	69.00	76.00	99.00	88.41	100.00	100.00	61.00	76.00	99.00
	2	56.00	67.00	95.00	73.21	100.00	100.00	41.00	67.00	95.00
3	1	38.00	77.00	84.00	18.42	74.03	100.00	7.00	57.00	84.00
	2	21.00	69.00	89.00	9.52	84.06	100.00	2.00	58.00	89.00

ROC: Rate of convergence, RAS: Rate of algorithm's success, ROSC: Rate of successful convergence
NM: Nelder-Mead, QN: Quasi-Newton, EM: Expectation-Maximisation

Table 3.2: Summary of the average parameter estimates and empirical standard errors of the acceptable fits over the 100 simulations for the NM, QN and EM algorithms.

No. of Phases	No. of Covs	Parameter	Sim	NM		QN		EM	
				Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.
2	1	γ_1	0.200	0.206	0.005	0.206	0.004	0.201	0.003
	2	γ_1	0.300	0.306	0.004	0.305	0.004	0.302	0.003
		γ_2	-0.400	-0.383	0.018	-0.409	0.013	-0.384	0.010
3	1	γ_1	0.600	0.605	0.015	0.603	0.003	0.598	0.002
	2	γ_1	0.600	0.580	0.410	0.602	0.003	0.601	0.002
		γ_2	-0.200	-0.198	0.140	0.217	0.010	-0.196	0.009

NM: Nelder-Mead, QN: Quasi-Newton, EM: Expectation-Maximisation
Acceptable fits: Those fits for which all parameters fall within an "acceptable range"; here considered to be between 0 and 10.
Est.: Mean parameter estimate, Std. Err.: Empirical standard error

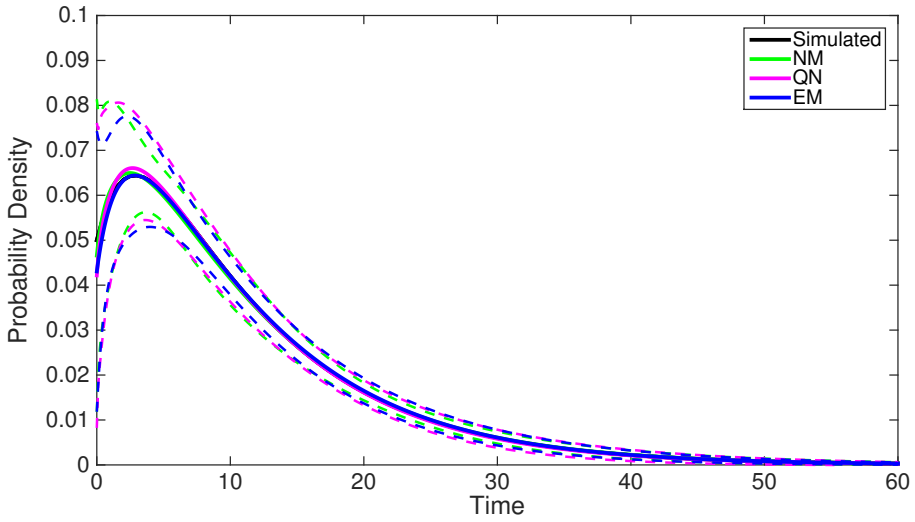


Figure 3.5: Mean baseline probability densities (and 95% confidence intervals) of the convergences of the NM, QN and EM algorithm approaches to fitting a three-phase Coxian with one covariate.

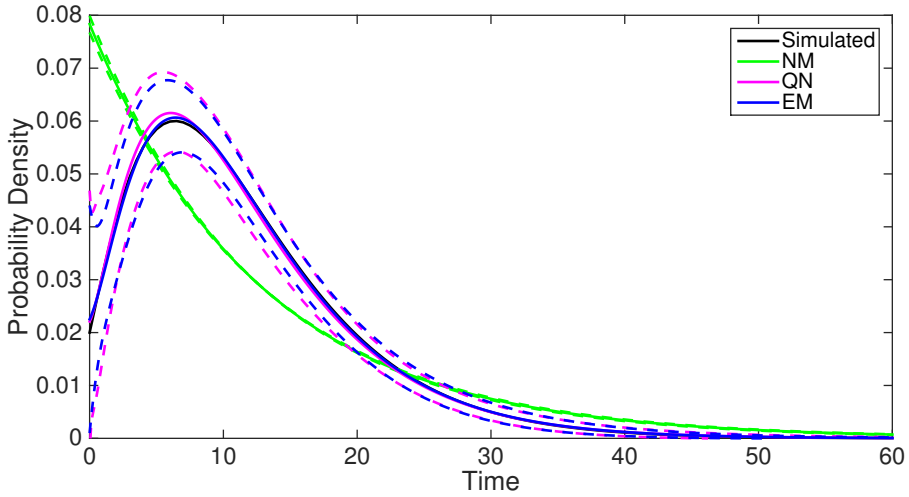


Figure 3.6: Mean baseline probability densities (and 95% confidence intervals) of the acceptable results of the NM, QN and EM algorithm approaches to fitting a three-phase Coxian with two covariates.

3.3.4 Summary

Step 1 Define starting values for the unknown parameters of the phase-type regression model: $\boldsymbol{\theta}_{\tau}^{(0)} = \{\mathbf{p}^{(0)}, \mathbf{T}^{(0)}, \boldsymbol{\gamma}^{(0)}\}$,

Step 2 Specify the latent variables which are to be approximated within the E-step of the algorithm, given by Equations 3.22 − 3.25,

Step 3 Iteratively work through the Expectation and Maximisation steps of the EM algorithm until a pre-defined convergence criteria is satisfied:

E-Step: Approximate the expected values of the latent variables using Equations 3.26 − 3.28,

M-Step: Differentiate the log-likelihood, given by Equation 3.21, with respect to the unknown parameters and solve to generate updated parameter estimates, given by Equations 3.29 − 3.31,

Step 4 Define the final parameter estimates based upon those from the final iteration of the algorithm.

3.4 Alternative Representation of the Covariate Effects

As previously discussed, the standard formulation of phase-type regression models imposes the restrictive assumption that a covariate will have a constant effect on all transitions through the underlying Markov process. Whilst this is done to reduce the number of additional parameters which are to be estimated by the already unstable fitting procedures found in current literature, it does so at the expense of limiting the information which can be gained regarding how a covariate impacts the system under investigation. In comparison, standard Markov modelling techniques allow for covariates to have a unique effect on each transition within the system.

A phase-type regression model which relaxes this assumption and instead allows for the effect of a covariate to vary across transitions offers a number of advantages, particularly in the cases of disease modelling. For example, many recent studies have been conducted to evaluate the effect of antiretroviral therapy on the survival of HIV patients, with a core interest in establishing when treatment intervention should commence so as to ensure the most significant impact. Kitahata et al. [156], for example, stratified their sample population based on the individuals' disease progression and were able to conclude that antiretroviral therapy had a much more significant effect when administered to patients in the early stages of HIV, and had a much reduced

impact on those in later stages. Fitting the standard phase-type regression model to such a scenario may cause the drug's significant impact on individuals' deterioration in the early stages of the disease to be masked by the insignificance in the later stages. By instead allowing the effect of the drug to vary amongst transitions, it would be possible to identify such a significant impact on only the earlier transitions through the system, thus informing intervention strategies.

Within this section, the increased stability of the EM algorithm approach to fitting phase-type regression models, developed within Section 3.3, is leveraged to relax this single covariate effect assumption by allowing the covariates to have varying effects across different transitions. It should be considered, however, that whilst the current formulation of the model may be too restrictive, a fully flexible model within which each covariate has a unique effect on each transition will introduce a large number of additional parameters which may be too liberal and computationally intensive to fit. To this end, three approaches of relaxing this limitation are considered:

i a state-specific (SS) approach, whereby a covariate is considered to have a single effect on all transitions from the same state, but varying effects across the states, as illustrated within Figure 3.7,

ii a direction-specific (DS) approach, whereby the effect of a covariate on the absorption transitions is considered to be different from the effect on the sequential transitions through the transient states, illustrated within Figure 3.8 and,

iii a transition specific (TS) approach, whereby the covariates have a unique effect on each transition within the system, illustrated within Figure 3.9.

These approaches are discussed in more detail within Sections 3.4.1 - 3.4.3 below and are validated by a simulation study in Section 3.4.4.

3.4.1 State Specific Parameterisation

In order to investigate the potential of a covariate's effect to vary depending upon the state from which the transition occurs, a state-specific (SS) parameterisation was developed within which the effect of the covariate is constant on all transitions from a single state, but can vary across the states, as illustrated for the Coxian within Figure 3.7. Within a Coxian phase-type regression model, where backwards transitions through the phases are not permitted, this parameterisation makes it possible to identify when a covariate has a significant impact. Under this formulation, the covariate parameter $\boldsymbol{\gamma}$ is now dependent upon state j, as shown:

$$q_{jk} = q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j\} \tag{3.36}$$

where $\boldsymbol{w}_i$ is a vector of time-invariant covariate parameters as before and q_{0jk} is the baseline transition intensity from state j into state k.

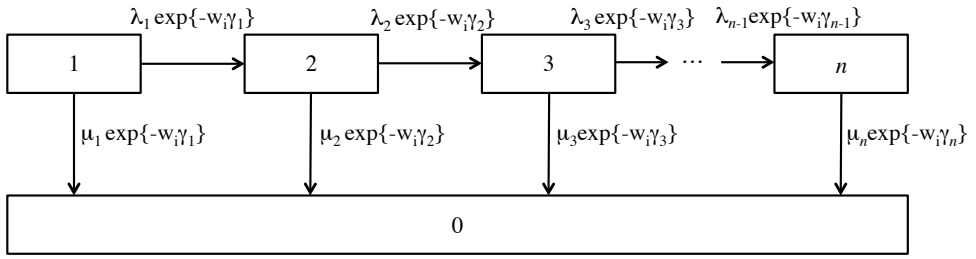


Figure 3.7: A Coxian phase-type distribution with state-specific covariate parameters, $\boldsymbol{\gamma}_j$.

The contribution to the overall likelihood of individual i is given by:

$$f_i(\tau_i; \boldsymbol{\theta}_\tau) = \prod_{j=1}^{n} p_j^{B_{ij}} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} \exp\left\{-q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j\} E_{ij}\right\} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} \left(q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j\}\right)^{N_{ijk}} \tag{3.37}$$

and the EM algorithm approach, developed within Section 3.3, is employed to maximise the likelihood, where the E-step remains largely unchanged with only the elements of the sub-generator matrix, $\mathbf{T}$, and exit vector, $\mathbf{t}$, updated to observe the new covariate parameterisation. Within the M-step, the baseline transition parameters are given by:

$$\widehat{q}_{0jk} = \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m} E_{ij} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j\}}, \qquad \text{for } j = 1, ..., n \ , \ k = 1, ..., n \text{ and } j \neq k \tag{3.38}$$

$$\widehat{q}_{0jj} = -\sum_{\substack{k=0 \\ k \neq j}}^{n} q_{0jk} \tag{3.39}$$

the initialisation vector remains unchanged and is given by Equation 3.30, and the

covariate parameters are again given by a one-step Newton Raphson process where the score vector, $S(\boldsymbol{\gamma}_j)$, and Hessian matrix, $H(\boldsymbol{\gamma}_j)$, are given by:

$$S(\boldsymbol{\gamma}_j) = \frac{\partial}{\partial \boldsymbol{\gamma}_j} \log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} q_{0jk} \boldsymbol{w}_i \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j\} E_{ij} - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} \boldsymbol{w}_i N_{ijk} \right\}$$

$$H(\boldsymbol{\gamma}_j) = \frac{\partial^2}{\partial {\boldsymbol{\gamma}_j}^2} \log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} -q_{0jk} \boldsymbol{w}_i \boldsymbol{w}_i' \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j\} E_{ij} \right\}. \tag{3.40}$$

3.4.2 Direction Specific Parameterisation

An alternative consideration is that a covariate may have a different effect on the event of interest occurring (i.e on the absorption transitions) compared to on the evolution of the disease (or transitions amongst the transient states of the system). For example, a treatment intervention may slow down a diseases progression but have an adverse effect on the individual's overall health and increase the hazard of absorption from each of the underlying states. For example, chemotherapy treatment in cancer patients may slow down the progression of the disease but could have a negative effect on other aspects of the individuals health, perhaps weakening the patient's immune system and increasing the risk of death from an associated comorbidity. Such a scenario can be evaluated by allowing each covariate to have two associated parameters: $\boldsymbol{\gamma}_0$, representing the effect of the covariate on the absorption rates, and $\boldsymbol{\gamma}_1$, representing the effect of the covariate on the rates of transition amongst the transient states:

$$\begin{aligned} q_{jk} &= q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_1\} \qquad k = 1, \ldots, n \\ q_{j0} &= q_{0j0} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_0\} \end{aligned} \tag{3.41}$$

This parameterisation is illustrated within Figure 3.8, and the contribution to the overall likelihood of individual i is given by:

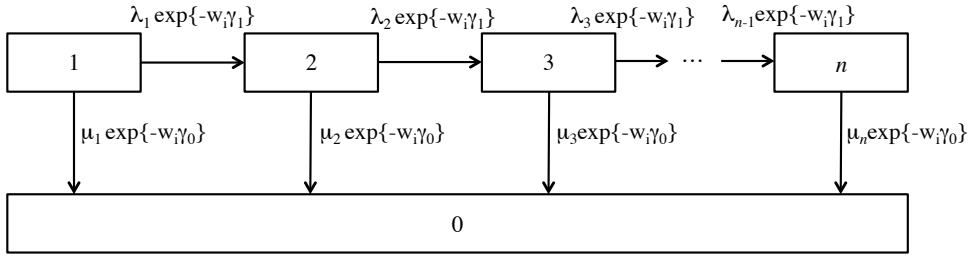


Figure 3.8: A Coxian phase-type distribution with direction-specific covariate parameters, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$.

$$f_i(\tau_i;\boldsymbol{\theta}_\tau) = \prod_{j=1}^{n} p_j^{B_{ij}} \prod_{j=1}^{n} \exp\Big\{ - q_{0j0} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_0\} E_{ij} \Big\} \prod_{j=1}^{n} \prod_{\substack{k=1 \\ k\neq j}}^{n} \exp\Big\{ - q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_1\} E_{ij} \Big\}$$
$$\times \prod_{j=1}^{n} \Big(q_{0j0} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_0\} \Big)^{N_{ij0}} \prod_{j=1}^{n} \prod_{\substack{k=1 \\ k\neq j}}^{n} \Big(q_{0jk} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_1\} \Big)^{N_{ijk}}. \tag{3.42}$$

Again, the E-step of the EM algorithm remains largely unchanged with just the elements of the sub-generator matrix, $\mathbf{T}$, and exit vector, $\mathbf{t}$, updated to observe the new covariate parameterisation, and within the M-step the transition intensities are given by:

$$\widehat{q}_{0jk} = \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m} E_{ij} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_1\}}, \qquad \text{for } j = 1, ..., n \ , \ k = 1, ..., n \text{ and } j \neq k \tag{3.43}$$

$$\widehat{q}_{0j0} = \frac{\sum_{i=1}^{m} N_{ij0}}{\sum_{i=1}^{m} E_{ij} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_0\}}, \tag{3.44}$$

$$\widehat{q}_{0jj} = -\sum_{\substack{k=0 \\ k\neq j}}^{n} \widehat{q}_{0jk} \tag{3.45}$$

the initialisation vector remains unchanged and is given by Equation 3.30, and the

covariate parameters are again given by a one-step Newton Raphson process where the score vectors, $S(\boldsymbol{\gamma}_0)$ and $S(\boldsymbol{\gamma}_1)$, are given by:

$$S(\boldsymbol{\gamma}_0) = \frac{\partial}{\partial \boldsymbol{\gamma}_0} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} q_{0j0} \boldsymbol{w}_i \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_0\} E_{ij} - \sum_{j=1}^{n} \boldsymbol{w}_i N_{ij0} \right\} \tag{3.46}$$

$$S(\boldsymbol{\gamma}_1) = \frac{\partial}{\partial \boldsymbol{\gamma}_1} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} q_{0jk} \boldsymbol{w}_i \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_1\} E_{ij} - \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \boldsymbol{w}_i N_{ijk} \right\} \tag{3.47}$$

and the Hessian matrices, $H(\boldsymbol{\gamma}_0)$ and $H(\boldsymbol{\gamma}_1)$, are given by:

$$\begin{aligned} H(\boldsymbol{\gamma}_0) &= \frac{\partial^2}{\partial {\boldsymbol{\gamma}_0}^2} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} -q_{0j0} \boldsymbol{w}_i \boldsymbol{w}_i' \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_0\} E_{ij} \right\} \\ H(\boldsymbol{\gamma}_1) &= \frac{\partial^2}{\partial {\boldsymbol{\gamma}_1}^2} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k \neq j}}^{n} -q_{0jk} \boldsymbol{w}_i \boldsymbol{w}_i' \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_1\} E_{ij} \right\}. \end{aligned} \tag{3.48}$$

3.4.3 Transition Specific Parameterisation

The final formulation to consider is a transition specific (TS) parameterisation whereby the covariates have a unique effect on each transition, in the same way as within a standard Markov model, as shown:

$$q_{jk} = q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_{jk}\} \tag{3.49}$$

where $\boldsymbol{\gamma}_{jk}$ is the effect of the covariate vector $\boldsymbol{w}_i$ on the transition from state j into state k.

This parameterisation is shown diagramatically within Figure 3.9 and the contribution to the likelihood of individual i is given by:

$$f_i(\tau_i; \boldsymbol{\theta}_\tau) = \prod_{j=1}^{n} p_j^{B_{ij}} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} \exp\Big\{ -q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_{jk}\} E_{ij} \Big\} \prod_{j=1}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{n} \Big(q_{0jk} \exp\{-\boldsymbol{w}_i' \boldsymbol{\gamma}_{jk}\} \Big)^{N_{ijk}}. \tag{3.50}$$

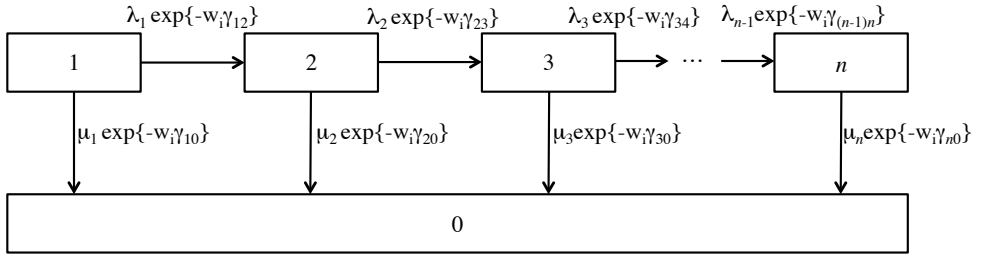


Figure 3.9: A Coxian phase-type distribution with transition-specific covariate parameters $\boldsymbol{\gamma}_{jk}$.

The E-step of the EM algorithm is updated so as the elements of the sub-generator matrix, $\mathbf{T}$, and exit vector $\mathbf{t}$, observe the new covariate parameterisation, as before. The updated transition parameters in the M-step are given by:

$$\begin{aligned} \widehat{q}_{0jk} &= \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m} E_{ij} \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_{jk}\}}, \qquad \text{for } j = 1, ..., n \ , \ k = 0, ..., n \text{ and } j \neq k \\ q_{0jj} &= -\sum_{\substack{k=0 \\ k\neq j}}^{n} \widehat{q}_{0jk} \end{aligned} \tag{3.51}$$

the initialisation vector remains unchanged and is given by Equation 3.30, and the covariate parameters are again given by a one-step Newton Raphson process where the score vector, $S(\boldsymbol{\gamma}_{jk})$, and Hessian matrix, $H(\boldsymbol{\gamma}_{jk})$ are given by:

$$S\big(\boldsymbol{\gamma}_{jk}\big) = \frac{\partial}{\partial \boldsymbol{\gamma}_{jk}} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk} \boldsymbol{w}_i \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_{jk}\} E_{ij} - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} \boldsymbol{w}_i N_{ijk} \right\}$$

$$H(\boldsymbol{\gamma}_{jk}) = \frac{\partial^2}{\partial {\boldsymbol{\gamma}_{jk}}^2} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \boldsymbol{w}_i \boldsymbol{w}_i' \exp\{-\boldsymbol{w}_i'\boldsymbol{\gamma}_{jk}\} E_{ij} \right\}.$$

3.4.4 Simulation Study Two

Within this simulation study, there are three primary targets of interest:

- Firstly, to validate the SS, DS and TS formulations of the new EM algorithm approach to fitting phase-type regression models,
- Secondly, to compare the performance of this new EM algorithm approach with the standard NM and QN algorithms previously employed, similarly adapted for the new covariate formulations,
- Finally, to investigate the loss of information when the constant effect formulation is applied to data which does not truly obey this assumption.

As within the simulation study conducted in Section 3.3.3, four scenarios were investigated whereby the number of underlying phases, as well as the number of covariates incorporated within the model, was increased to assess how well the different algorithms handled an increasing number of unknown parameters. To this end, for each of the covariate approaches under investigation, four datasets were simulated to be Coxian phase-type distributed using the actuar package in R [153], consisting of:

i two underlying phases, defined by baseline transition parameters $q_{010} = 0.05$, $q_{020} = 0.10$ and $q_{012} = 0.30$, where the effect of the single continuous covariate, $w_{i1} \sim \text{unif}(-3, 3)$, on the transition intensities for each of the three approaches under investigation is given by:

- SS parameterisation: $\boldsymbol{\gamma}_1 = 0.2$, $\boldsymbol{\gamma}_2 = -0.4$
- DS parameterisation: $\boldsymbol{\gamma}_0 = -0.5$, $\boldsymbol{\gamma}_1 = 0.4$
- TS parameterisation: $\boldsymbol{\gamma}_{10} = 0.3$, $\boldsymbol{\gamma}_{20} = 0.7$, $\boldsymbol{\gamma}_{12} = -0.5$

ii two underlying phases, defined by baseline transition parameters $q_{010} = 0.05$, $q_{020} = 0.10$ and $q_{012} = 0.30$, where the effect of the two continuous covariates, $w_{i1} \sim \text{unif}(-3, 3)$ and $w_{i2} \sim \text{unif}(-3, 3)$, on the transition intensities for each of the three approaches under investigation is given by:

- SS parameterisation: $\boldsymbol{\gamma}_1 = \begin{pmatrix} 0.1 \\ -0.3 \end{pmatrix}$, $\boldsymbol{\gamma}_2 = \begin{pmatrix} -0.6 \\ 0.4 \end{pmatrix}$
- DS parameterisation: $\boldsymbol{\gamma}_0 = \begin{pmatrix} 0.7 \\ -0.2 \end{pmatrix}$, $\boldsymbol{\gamma}_1 = \begin{pmatrix} 0.1 \\ -0.3 \end{pmatrix}$
- TS parameterisation: $\boldsymbol{\gamma}_{10} = \begin{pmatrix} 0.1 \\ -0.6 \end{pmatrix}$, $\boldsymbol{\gamma}_{20} = \begin{pmatrix} 0.5 \\ -0.2 \end{pmatrix}$, $\boldsymbol{\gamma}_{12} = \begin{pmatrix} 0.3 \\ -0.3 \end{pmatrix}$

iii three underlying phases, defined by baseline transition parameters $q_{010} = 0.02$, $q_{020} = 0.04$, $q_{030} = 0.06$, $q_{012} = 0.10$ and $q_{023} = 0.20$, where the effect of the single

continuous covariates, $w_{i1} \sim \text{unif}(-3, 3)$, on the transition intensities for each of the three approaches under investigation is given by:

- SS parameterisation: $\boldsymbol{\gamma}_1 = 0.5$, $\boldsymbol{\gamma}_2 = -0.4$, $\boldsymbol{\gamma}_3 = 0.8$
- DS parameterisation: $\boldsymbol{\gamma}_0 = -0.3$, $\boldsymbol{\gamma}_1 = 0.2$
- TS parameterisation: $\boldsymbol{\gamma}_{10} = -0.1$, $\boldsymbol{\gamma}_{20} = -0.3$, $\boldsymbol{\gamma}_{30} = -0.4$, $\boldsymbol{\gamma}_{12} = 0.2$, $\boldsymbol{\gamma}_{23} = 0.4$

iv three underlying phases, defined by baseline transition parameters $q_{010} = 0.02$, $q_{020} = 0.04$, $q_{030} = 0.06$, $q_{012} = 0.10$ and $q_{023} = 0.20$, where the effect of the two continuous covariates, $w_{i1} \sim \text{unif}(-3, 3)$ and $w_{i2} \sim \text{unif}(-2, 2)$, on the transition intensities for each of the three approaches under investigation is given by:

- SS parameterisation: $\boldsymbol{\gamma}_1 = \begin{pmatrix} 0.15 \\ -0.25 \end{pmatrix}$, $\boldsymbol{\gamma}_2 = \begin{pmatrix} -0.1 \\ 0.3 \end{pmatrix}$, $\boldsymbol{\gamma}_3 = \begin{pmatrix} 0.3 \\ -0.2 \end{pmatrix}$
- DS parameterisation: $\boldsymbol{\gamma}_0 = \begin{pmatrix} 0.2 \\ -0.05 \end{pmatrix}$, $\boldsymbol{\gamma}_1 = \begin{pmatrix} 0.1 \\ -0.3 \end{pmatrix}$
- TS parameterisation: $\boldsymbol{\gamma}_{10} = \begin{pmatrix} -0.30 \\ 0.30 \end{pmatrix}$, $\boldsymbol{\gamma}_{20} = \begin{pmatrix} -0.2 \\ 0.1 \end{pmatrix}$, $\boldsymbol{\gamma}_{30} = \begin{pmatrix} -0.40 \\ 0.20 \end{pmatrix}$, $\boldsymbol{\gamma}_{12} = \begin{pmatrix} 0.50 \\ -0.20 \end{pmatrix}$, $\boldsymbol{\gamma}_{23} = \begin{pmatrix} 0.10 \\ -0.50 \end{pmatrix}$

Within this study, only one dataset was simulated for each of the approaches under investigation, as opposed 100 different datasets, as was the case for Simulation Study One. The reason for this change in procedure is that this simulation study aims to investigate the impact of the starting values of the unknown parameters on the performance of the algorithms under investigation. Using the same dataset for 100 different initialisations, where only the parameter starting values vary across initialisation, means that any differences observed can be attributed solely to the starting values, and not to differences in the datasets used.

The datasets for each approach were simulated to consist of more observations than previously, 1500 as opposed to 400, so as to minimise sample bias. To ensure that each algorithm yielded enough convergences to analyse, enough initialisations were performed on each algorithm so as to generate 100 successful convergences, instead of performing just 100 initialisations as was done within Simulation One. The RAS performance measure from the first simulation was retained, and two new measures were introduced, based upon the mean relative distance to the maximum likelihood estimates (MRD) measure, previously employed within the literature [138, 154]. The MRD, along with the two adjustments that were employed here, are described below:

a. Mean relative distance to the MLEs (MRD)

This is a measure of how close the estimated parameter values are to the true simulated values, where the MRD for the l^{th} successful convergence is given by [145]:

$$\text{MRD}_l = \frac{\sum_{w=1}^{W} \frac{|\zeta_w - \widehat{\zeta}_w|}{\zeta_w}}{\text{number of parameters}} \tag{3.52}$$

where W is the number of estimated parameters, ζ_w is the true parameter value and $\widehat{\zeta}_w$ is the value estimated by the l^{th} convergence.

This value can then be averaged over the s convergences, as shown:

$$\text{MRD} = \frac{\sum_{l=1}^{s} \text{MRD}_l}{\text{number of convergences}}. \tag{3.53}$$

b. MRD of acceptable fits (MRDa)

Instead of calculating the MRD for all convergences, the MRD was instead calculated for only those fits which were deemed "acceptable". Different algorithms may be more prone to converging to local maxima, which results in an "unacceptable result", therefore introducing bias to the MRD. However, the RAS metric already identifies to what extent this occurs for each of the approaches. Therefore, the MRDa controls for this bias by evaluating only the quality of those fits from each algorithm which are deemed acceptable, and is given by:

$$\text{MRDa} = \frac{\sum_{q=1}^{A} \text{MRDa}_q}{\text{number of acceptable results}} \tag{3.54}$$

where A is the total number of acceptable results and MRDa_q is the mean relative distance calculated using Equation 3.52 for the q^{th} acceptable result.

c. MRD of covariate parameters (MRDc)

As mentioned previously, phase-type distributions suffer from a non-singularity problem whereby the same shape distribution can be represented by multiple sets of parameters. This means that it is possible to accurately represent the shape of the distribution with a set of parameters which are different from the simulated values, resulting in a high MRDa value, despite the fact that they suitably

represent the data. To remove this effect, and instead evaluate the MRD of only the covariate parameters, the MRDc measure is introduced. The contribution to the MRDc of the q^{th} acceptable result can be given by Equation 3.52, where ζ_w represents the w^{th} covariate parameter. The MRDc is then calculated by:

$$\text{MRDc} = \frac{\sum_{q=1}^{A} MRDc_q}{\text{number of acceptable results}} \tag{3.55}$$

3.4.4.1 Results

The RAS, MRDa and MRDc scores for the four approaches are given in Table 3.3. Looking first a the RAS scores, it can be observed that the NM produces the fewest acceptable results for each of the SS, DS and TS approaches, just as was observed for the constant effect (CE) approach within Simulation One, with the EM algorithm again performing best in terms of the RAS. Similarly, the MRDa and MRDc scores also indicate the EM algorithm to be the superior of the three fitting procedures for each of the approaches, as the scores for the EM algorithm are, in general, closest to zero. It can be seen that as the number of unknown parameters increases, both the MRDa and MRDc scores increase, indicating that the algorithms begin to increasingly suffer from the aforementioned identifiability issues.

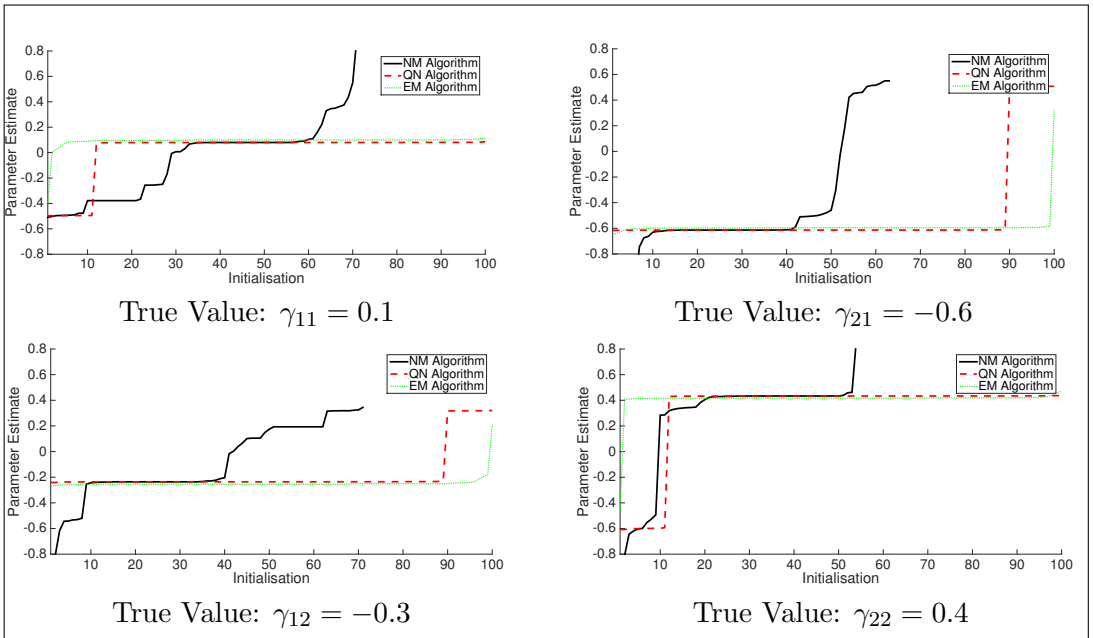
The higher MRDc scores within Table 3.3 suggest that not all acceptable fits successfully estimate the true covariate parameter values. Looking more closely at the SS scenario, the estimated covariate parameters from each of the acceptable fits for the two-phase simulation, for example, are plotted within Figure 3.10. It can be seen that the EM algorithm has the highest success rate, as it has the most fits which estimate the true values of the parameters, which is further validated by the EM algorithm having the lowest MRDc score of 0.091, compared to 3.302 and 0.457 for the NM and QN algorithms, respectively.

Table 3.3: Summary of the RAS, MRDa and MRDc scores over the 100 simulations for the NM, QN and EM algorithms for the four approaches to incorporating covariates within the Coxian phase-type distribution.

Approach	**No of Phases**	**No of Covariates**	**RAS**			**MRDa (MRDc)**		
			NM	**QN**	**EM**	**NM**	**QN**	**EM**
SS	2	1	91.00	93.00	100.00	1.034 (1.055)	0.269 (0.215)	0.234 (0.145)
		2	71.00	100.00	100.00	2.543 (3.302)	0.511 (0.457)	0.126 (0.091)
	3	1	35.00	80.00	100.00	4.992 (5.393)	2.864 (1.094)	0.874 (0.944)
		2	36.00	91.00	100.00	10.13 (13.19)	1.830 (1.041)	1.605 (1.262)
DS	2	1	73.00	100.00	100.00	1.925 (3.463)	0.718 (0.556)	0.684 (0.610)
		2	82.00	100.00	100.00	13.817 (23.079)	0.588 (0.712)	1.018 (1.200)
	3	1	45.00	96.00	96.00	4.042 (1.290)	3.183 (1.348)	1.881 (0.403)
		2	39.00	92.00	100.00	4.749 (5.293)	2.400 (0.941)	1.416 (1.021)
TS	2	1	81.00	100.00	100.00	3.346 (5.778)	1.405 (1.101)	0.438 (0.400)
		2	65.00	100.00	100.00	6.680 (9.139)	1.722 (1.639)	0.544 (0.397)
	3	1	38.00	90.00	100.00	7.245 (10.274)	2.989 (1.975)	1.340 (1.279)
		2	46.00	79.00	100.00	6.873 (7.982)	3.114 (2.632)	1.842 (2.000)

RAS: Rate of algorithm's success, MRDa: Mean relative distance of all unknown paramaters utilising the acceptable fits, MRDc: Mean relative distance of all unknown covariate parameters utilising the acceptable fits
NM: Nelder-Mead algorithm, QN: Quasi-Newton algorithm, EM: Expectation-Maximisation algorithm
SS: State-specific approach, DS: Direction specific approach, TS: Transition specific approach

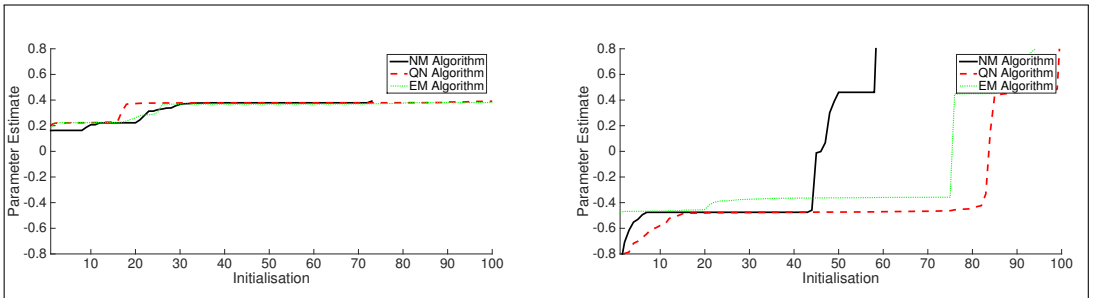
For the DS scenario, it can be observed, once again, that not all of the acceptable fits resulted in an accurate estimate of the 'true' simulated covariate parameter value. Figure 3.11, for example, shows the results of the two phase simulation with one covariate where it can be observed, similarly to before, that the EM algorithm provides more accurate estimations of the true effect than the NM and QN.



True Value: $\gamma_{11} = 0.1$ True Value: $\gamma_{21} = -0.6$

True Value: $\gamma_{12} = -0.3$ True Value: $\gamma_{22} = 0.4$

Figure 3.10: Comparison of the ordered covariate parameter values from the 100 simulations, estimated by employing the NM, QN and EM algorithm approaches to data simulated according to a two phase Coxian distribution with two covariates, observing the SS parameterisation.



True Value: $\gamma_1 = 0.4$ True Value: $\gamma_0 = -0.5$

Figure 3.11: Comparison of the ordered covariate parameter values from the 100 simulations, estimated by employing the NM, QN and EM algorithm approaches to data simulated according to a two phase Coxian distribution with one covariate, observing the DS parameterisation.

Finally, for the TS scenario it can be inferred from the overall higher MRDc values that fewer of the acceptable fits result in successful estimations of the true covariate effects, particularly for the NM algorithm. This is shown visually, for example, for the two phase simulation with one covariate in Figure 3.12, where it can be observed that the EM algorithm once again proved to generate the most successful estimations of the covariate effect.

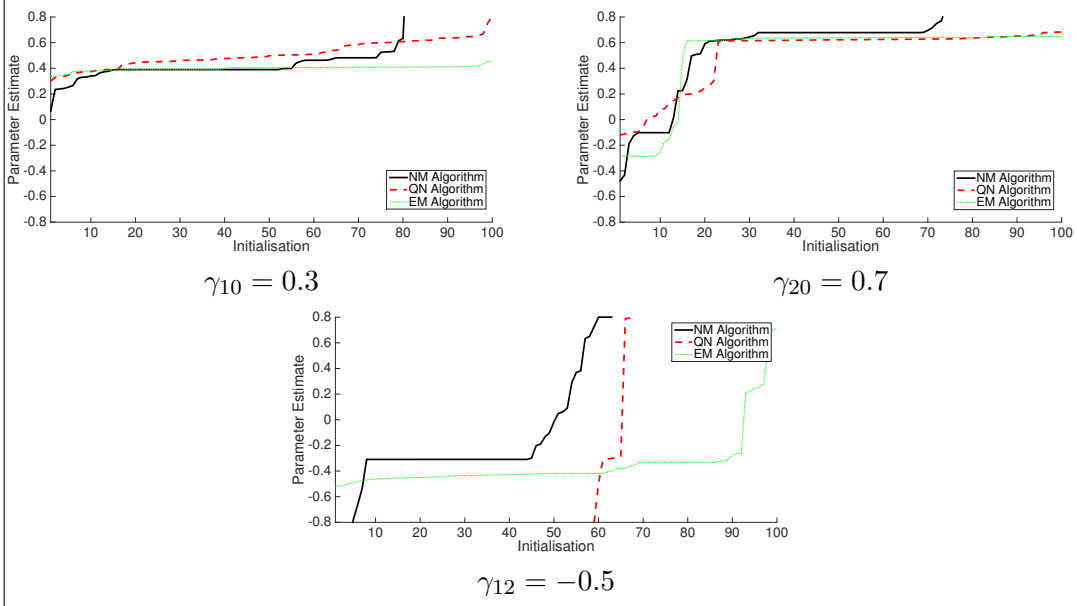


Figure 3.12: Comparison of the ordered covariate parameter values from the 100 simulations, estimated by employing the NM, QN and EM algorithm approaches to data simulated according to a two phase Coxian distribution with one covariate, observing the TS parameterisation.

The final aim of the simulation study is to investigate how successful the constant effect (CE) approach is at representing the behaviour of the model when the true effect of the covariate is not constant across all transitions. In order to investigate this, the standard approach which assumes a constant covariate effect was used to fit a phase-type distribution to the four datasets simulated to follow the TS approach. The true simulated values of the covariates on each transition rate, along with the estimated covariate effects when using the CE approach, are given in Table 3.4.

Table 3.4: Table showing the estimated covariate parameters when a Coxian phase-type regression model which assumes a CE approach is fitted to data simulated according to a TS approach.

Simulated Parameter Values Under the TS Assumption:				
Two Phases			**Three Phases**	
One Covariate	**Two Covariates**		**One Covariate**	**Two Covariates**	
$\gamma_{110} = 0.300$	$\gamma_{110} = 0.100$	$\gamma_{210} = -0.600$	$\gamma_{110} = -0.100$	$\gamma_{110} = -0.300$	$\gamma_{210} = 0.300$
$\gamma_{120} = 0.700$	$\gamma_{120} = 0.500$	$\gamma_{220} = -0.200$	$\gamma_{120} = -0.300$	$\gamma_{120} = -0.200$	$\gamma_{220} = 0.100$
$\gamma_{112} = -0.500$	$\gamma_{112} = 0.300$	$\gamma_{212} = -0.300$	$\gamma_{130} = -0.400$	$\gamma_{130} = -0.400$	$\gamma_{230} = 0.200$
			$\gamma_{112} = 0.200$	$\gamma_{112} = 0.500$	$\gamma_{212} = -0.200$
			$\gamma_{123} = 0.400$	$\gamma_{123} = 0.100$	$\gamma_{223} = -0.500$
Estimated Parameter Values Assuming a CE Approach:				
$\gamma_{1\cdot\cdot} = 0.461$	$\gamma_{1\cdot\cdot} = 0.416$	$\gamma_{2\cdot\cdot} = -0.285$	$\gamma_{1\cdot\cdot} = -0.175$	$\gamma_{1\cdot\cdot} = -0.125$	$\gamma_{2\cdot\cdot} = 0.044$

TS: Transition specific, CE: Constant effect

It can be observed that the estimated effect by the CE approach serves as an almost pseudo-average of the overall effect and so, in cases where the covariate has a similar effect on all transitions it may be sufficient to employ the CE approach. For example, for the two phase simulation with two covariates it can be observed that the true effect of the first covariate on each transition is always positive; $\gamma_{110} = 0.100$, $\gamma_{120} = 0.500$ and $\gamma_{112} = 0.300$, which is reflected in the estimated effect under the constant effect assumption, $\gamma_{1\cdot\cdot} = 0.416$.

However, in cases where the effect of the covariate varies more extremely across the transitions, the CE estimate may not provide information which reflects the true effect of the covariate. For example, for the three phase simulation with two covariates, the second covariate has a positive effect on each of the absorption transitions; $\gamma_{210} = 0.300$, $\gamma_{220} = 0.100$ and $\gamma_{230} = 0.200$, and a negative effect on the sequential transitions; $\gamma_{212} = -0.200$ and $\gamma_{213} = -0.500$. The estimated 'overall' effect by the CE approach is 0.044, which is much weaker than the true transition-specific effects, giving the false impression that the covariate is not strongly associated with the individuals rates of flow through the system.

3.5 Inhomogeneous Coxian Phase-type Regression Model

Within standard Coxian phase-type literature, the transition intensities which define the distribution are assumed to be constant over time, reflecting the assumption that the underlying Markov process is time homogeneous. However, if interest lies in incorporating a time-varying covariate within a Coxian phase-type regression model, such as a longitudinal response, this homogeneity assumption is violated as the transition intensity now varies with time, as shown:

$$q_{jk}(t) = q_{0jk} \exp\big\{ - w_i(t)\gamma \big\} \tag{3.56}$$

where $w_i(t)$ is an observation on a covariate value at time t and q_{0jk} is the baseline transition intensity, representing the rate of transition from state j into state k when $w_i(t) = 0$.

Within this section, the EM algorithm approach to fitting phase-type regression models, developed within Section 3.3, is further extended to allow for the incorporation of time-varying covariates within the specific case of the Coxian, something which has never previously been explored within the literature. Doing so significantly extends the scope of Coxian phase-type regression models, particularly within medical statistics, where longitudinal biomarkers and time-varying endogenous covariates are

regularly of interest within time-to-event studies. Before adapting the density function and discussing the fitting procedure, let us conceptually discuss the progress of an individual through an n-phase Coxian distribution when the transition intensities are influenced by a time-varying covariate.

At time zero, individual i belongs to the first state of the underlying Markov process, where their initial rates of transitioning into the absorbing and second state of the process are given by:

$$h_{i10}\big(0\big) = q_{010} \exp\big\{ - w_i(0)\gamma\big\} \tag{3.57}$$
$$h_{i12}\big(0\big) = q_{012} \exp\big\{ - w_i(0)\gamma\big\} \tag{3.58}$$

If E_{i1} is the expected time individual i spends in state one, the rate of transitioning into the absorbing and second state at the expected transition time are given by:

$$h_{i10}\big(E_{i1}\big) = q_{010} \exp\big\{ - w_i(E_{i1})\gamma\big\} \tag{3.59}$$
$$h_{i12}\big(E_{i1}\big) = q_{012} \exp\big\{ - w_i(E_{i1})\gamma\big\} \tag{3.60}$$

where the corresponding survivor functions are:

$$S_{i10}\big(E_{i1}\big) = \exp\left\{ - \int_0^{E_{i1}} q_{010} \exp\big\{ - w_i(s)\gamma\big\} ds \right\} \tag{3.61}$$
$$S_{i12}\big(E_{i1}\big) = \exp\left\{ - \int_0^{E_{i1}} q_{012} \exp\big\{ - w_i(s)\gamma\big\} ds \right\}. \tag{3.62}$$

If individual i makes the transition into the second state at time E_{i1}, their rate of transitioning into either the absorbing or third states at time E_{i1} are given by:

$$h_{i20}\big(E_{i1}\big) = q_{020} \exp\big\{ - w_i(E_{i1})\gamma\big\} \tag{3.63}$$
$$h_{i23}\big(E_{i1}\big) = q_{023} \exp\big\{ - w_i(E_{i1})\gamma\big\} \tag{3.64}$$

and if E_{i2} is the expected time spent within this second state, the rates of transitioning into either the absorbing or third states at their expected transition time are given by:

$$h_{i20}\big(E_{i1}+E_{i2}\big) = q_{020}\exp\big\{-w_i(E_{i1}+E_{i2})\gamma\big\} \tag{3.65}$$

$$h_{i23}\big(E_{i1}+E_{i2}\big) = q_{023}\exp\big\{-w_i(E_{i1}+E_{i2})\gamma\big\} \tag{3.66}$$

with corresponding survivor functions:

$$S_{i20}\big(E_{i1}+E_{i2}\big) = \exp\left\{-\int_{E_{i1}}^{E_{i1}+E_{i2}} q_{020}\exp\big\{-w_i(s)\gamma\big\}ds\right\} \tag{3.67}$$

$$S_{i23}\big(E_{i1}+E_{i2}\big) = \exp\left\{-\int_{E_{i1}}^{E_{i1}+E_{i2}} q_{023}\exp\big\{-w_i(s)\gamma\big\}ds\right\} \tag{3.68}$$

and so on until the n^{th} state from which the individual can only absorb.

The probability density function of this time inhomogeneous Markov process is thus given by:

$$\begin{aligned} f(\tau_i;\boldsymbol{\theta}_\tau) = &\prod_{j=1}^{n} p_j^{B_{ij}} \prod_{j=1}^{n}\prod_{\substack{k=0\\k\neq j}}^{n} \exp\left\{-\int_{F_{ij}}^{F_{i(j+1)}} q_{0jk}\exp\big\{-w_i(s)\gamma\big\}ds\right\} \\ &\times \prod_{j=1}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\Big(q_{0jk}\exp\big\{-w_i(F_{i(j+1)})\gamma\big\}\Big)^{N_{ijk}} \end{aligned} \tag{3.69}$$

where F_i is a vector denoting the cumulative time spent within each state for individual i, where $F_{ij} = \sum_{j*=1}^{j-1} E_{ij^*}$, where $j \neq 1$ and $F_{i1} = 0$, i.e:

$$F_i = \Big(0,\quad E_{i1},\quad E_{i1}+E_{12},\quad E_{i1}+E_{12}+E_{13},\ \ldots\Big). \tag{3.70}$$

The corresponding log-likelihood is given by:

$$\log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} B_{ij} \log(p_j) - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} \int_{F_{ij}}^{F_{i(j+1)}} q_{0jk} \exp\big\{ - w_i(s)\gamma \big\} ds + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} N_{ijk} \Big(\log(q_{0jk}) - w_i(F_{i(j+1)})\gamma \Big) \Bigg\}. \tag{3.71}$$

3.5.1 E-Step

Within the E-Step of the time inhomogeneous Coxian phase-type regression model it is necessary to approximate N_{ijk} and N_{ij0} as before, where the time-varying nature of the covariate must be taken into consideration. However, instead of approximating E_{ij}, it is now necessary to approximate $\int_{F_{ij}}^{F_{i(j+1)}} \exp\big\{ - w_i(s)\gamma \big\} ds$, along with the first and second derivatives with respect to γ. As such, the expected values of the latent variables on any iteration of the EM algorithm are given by:

$$\mathbf{E}\Big[N_{ijk} \mid \tau_i \Big] = \frac{q_{0jk} c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.72}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} \exp\big\{ - w_i(s)\boldsymbol{\gamma} \big\} ds \mid \tau_i \right] = \frac{c_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.73}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} w_i(s) \exp\big\{ - w_i(s)\boldsymbol{\gamma} \big\} ds \mid \tau_i \right] = \frac{g2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.74}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} w_i(s) w_i(s)' \exp\big\{ - w_i(s)\boldsymbol{\gamma} \big\} ds \mid \tau_i \right] = \frac{g3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.75}$$

and, for individuals who experience the event of interest:

$$\mathbf{E}\Big[N_{ij0} \mid \tau_i \Big] = \frac{a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) q_{0j0} \exp\{-w_i(\tau_i)\gamma\}}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{3.76}$$

where $N_{ij0} = 0$ for censored individuals. The values of $a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $g1_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $g2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)$ and $g3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)$ are given by:

$$a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{\tau_i} \exp\{ - w_i(s)\gamma \} ds \right\} \mathbf{e_j} \tag{3.77}$$

$$d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{e_j}' \exp\left\{ \mathbf{T} \int_0^{\tau_i} \exp\{ - w_i(s)\gamma \} ds \right\} \left(\mathbf{t} \exp\{ - w_i(\tau_i)\gamma \} \right)^{\delta_i} \tag{3.78}$$

$$\begin{aligned} c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{s} \exp\{ - w_i(u)\gamma \} du \right\} \mathbf{e_j} \exp\{ - w_i(s)\gamma \} \\ \mathbf{e_k}' \exp\left\{ \mathbf{T} \int_s^{\tau_i} \exp\{ - w_i(u)\gamma \} du \right\} \left(\mathbf{t} \exp\{ - w_i(\tau_i)\gamma \} \right)^{\delta_i} ds \end{aligned} \tag{3.79}$$

$$\begin{aligned} g2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{s} \exp\{ - w_i(u)\gamma \} du \right\} \mathbf{e_j} w_i(s) \exp\{ - w_i(s)\gamma \} \\ \mathbf{e_j}' \exp\left\{ \mathbf{T} \int_s^{\tau_i} \exp\{ - w_i(u)\gamma \} du \right\} \left(\mathbf{t} \exp\{ - w_i(\tau_i)\gamma \} \right)^{\delta_i} ds \end{aligned} \tag{3.80}$$

$$\begin{aligned} g3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{s} \exp\{ - w_i(u)\gamma \} du \right\} \mathbf{e_j} w_i(s) w_i(s)' \exp\{ - w_i(s)\gamma \} \\ \mathbf{e_j}' \exp\left\{ \mathbf{T} \int_s^{\tau_i} \exp\{ - w_i(u)\gamma \} du \right\} \left(\mathbf{t} \exp\{ - w_i(\tau_i)\gamma \} \right)^{\delta_i} ds \end{aligned} \tag{3.81}$$

3.5.2 M-Step

Within the M-step of the algorithm, updated estimates for the unknown transition intensities are given by:

$$\widehat{q}_{0jk} = \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m} \int_{F_{ij}}^{F_{i(j+1)}} \exp\{-w_i(s)\gamma\} ds}, \quad \text{for } j = 1, ..., n \ , \ k = 0, ..., n \text{ and } j \neq k \tag{3.82}$$

$$\widehat{q}_{0jj} = -\sum_{\substack{k=0 \\ k \neq j}}^{n} q_{0jk} \tag{3.83}$$

A one-step Newton Raphson process is again used to obtain estimates of the co-

variate parameters, γ:

$$\widehat{\gamma}^{(l+1)} = \widehat{\gamma}^{(l)} - H\big(\widehat{\gamma}^{(l)}\big)^{-1} S\big(\gamma^{(l)}\big) \tag{3.84}$$

where $S\big(\gamma\big)$ is the score vector, given by:

$$S(\gamma) = \frac{\partial}{\partial \gamma} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} w_i(s) \exp\{-w_i(s)\gamma\} ds - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} w_i(F_{i(j+1)}) N_{ijk} \right\} \tag{3.85}$$

and $H\big(\gamma\big)$ is the Hessian, given by:

$$H(\gamma) = \frac{\partial^2}{\partial \gamma^2} \log L\big(\boldsymbol{\theta}_\tau; \tau_i\big) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k \neq j}}^{n} -q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} w_i(s) w_i(s)' \exp\{-w_i(s)\gamma\} ds \right\}. \tag{3.86}$$

3.6 Summary

This chapter began with a review of the methodology and current applications of both phase-type distributions and phase-type regression models. Through this, a number of features of the Coxian phase-type regression model which make it an attractive approach to represent survival data within a joint modelling framework were highlighted. In particular:

i through the latent phases uncovered during the fitting process, the Coxian phase-type distribution provides further insight into the failure process which it represents, in comparison to standard survival models. By mapping these uncovered states onto distinct stages of the survival process, inferences can be made regarding the rates of deterioration through these stages and, thus, the quality of life individuals will experience for their remaining survival time.

ii the Coxian phase-type distribution is capable of representing any positive distribution to an arbitrary degree of accuracy, overcoming the limitations associated

with alternative parametric survival models which are limited in terms of the distributional shapes which they can suitably fit [125],

iii Coxian phase-type regression models are fully parametric, alleviating previously identified limitations which are encountered when a semi-parametric survival model is incorporated within a joint model, resulting in the underestimation of the standard errors [11],

iv the fully parametric nature of the Coxian phase-type regression model is also advantageous compared to semi-parametric approaches in terms of making survival predictions from the fitted joint model [12].

Despite their advantages, however, there are some limitations associated with phase-type regression models, well documented within the literature. A number of these limitations were highlighted within this research, and new methodology was developed to overcome them, improving the suitability of phase-type distributions for representing the survival process within a joint model. Specifically:

a. a new EM algorithm approach to fitting phase-type regression models was developed, and shown to be more stable, both in terms of its rate of successful convergence and in its precision when estimating covariate effects, compared to previously employed NM and QN algorithm approaches,

b. this new EM algorithm approach, evidenced through its high RAS scores, also showed a significant reduction in the number of false convergences, well documented to plague the fitting of phase-type distributions through alternative approaches, meaning excessive initialisations are no longer necessary as significantly more successful convergences resulted in an acceptable fit,

c. the increased stability observed within this newly developed EM algorithm approach was leveraged to relax the restrictive assumption that covariates have a constant effect on all transitions through the underlying Markov model. Instead, new formulations of the model were presented which allow state-specific, direction-specific and transition-specific inferences to be drawn from the data, where the NM and QN were shown to struggle with these adaptions,

d. the new EM algorithm approach was also extended to allow for the incorporation of time-varying covariates, not previously investigated within phase-type literature, broadening the scope and applicability of the models.

3.6. Summary

Within Chapter 4, new methodology for incorporating the Coxian phase-type regression model within a joint likelihood, alongside a LME model, is developed. This novel joint model is subsequently applied to a dataset collected on individuals suffering from chronic kidney disease within Chapter 5.

Chapter 4

Joint Modelling of Longitudinal and Survival Data Utilising the Coxian Phase-type Distribution

4.1 Overview

This chapter details the development of a new joint modelling approach to the analysis of longitudinal and survival data, within which the Coxian phase-type regression model is employed to represent the survival process. This new joint model observes the standard shared parameter formulation, found within the literature, where the association between the longitudinal and survival processes is represented by latent random effects.

Preliminarily, a two-stage approach is explored, similar to that of the early two stage procedures proposed by Tsiatis et al. [5] and Self and Pawitan [6] to incorporate unbiased estimates of a repeatedly observed covariate within a survival model. Subsequently, the joint likelihood approach to this new methodology is detailed, where two common joint modelling parameterisations are developed:

i the random effects (RE) parameterisation, where the individuals' latent random effects are incorporated within the survival submodel, and,

ii the true longitudinal response (TLR) parameterisation, where an unbiased prediction of the individual's longitudinal response is incorporated within the survival submodel.

Finally, a simulation study is conducted both to validate the new methodology and to compare the approach with the commonly employed exponential and Weibull AFT joint models, available through the JM package within R software [14].

4.2 Motivation

Employing the Coxian phase-type distribution to represent the survival process within a joint modelling framework offers a number of advantages to both these previously independent areas of statistical research, broadening their applicability and scope. Previously, Coxian phase-type distributions have been utilised extensively to model patient flow through hospital, favoured for their ability to uncover underlying stages of the process which they represent [25, 26, 131, 132, 139, 140]. They have also been employed, although somewhat less commonly, to model heavy tailed distributions [157–159], here valued for their ability to represent any positive distribution to an arbitrary degree of accuracy, and they have been applied within the area of risk theory to estimate ruin probabilities, favoured for their mathematical tractability [160]. The utilisation of phase-type distributions within more typical survival analysis-type problems, on

the other hand, remains novel, with limited investigation beyond that of Aalen [28], who first discussed their possible applicability to biostatistics survival problems. The incorporation of covariates within such models, allowing their effect on the tranitiion intensity parameters to be evaluated, is similarly under-investigated, with only a limited number of techniques discussed within the literature [30, 147, 148]. Further, in order to aid successful convergence, these covariate approaches impose the restrictive assumption that a covariate will have the same effect across all transition intensities, limiting the depth of the information which can be ascertained regarding how covariates affect the system under investigation.

The first stage of this research, detailed within Chapter 3, targeted some of the more impeding limitations of phase-type regression models, with the aim of improving their suitability for representing the survival process within a joint modelling framework. Specifically, a newly developed EM algorithm approach to fitting phase type-regression models was developed, detailed within Section 3.3, improving both the rate of successful convergence of the model, along with the accuracy of the parameter estimates themselves, when compared to previous estimation approaches, as illustrated through a simulation study within Section 3.3.3. Additionally, the restrictive assumption that covariates have a constant effect across all transitions was relaxed within this new EM algorithm approach, discussed within Section 3.4, increasing the level of insight which can be obtained pertaining to how the covariates affect the system represented by the phase-type distribution.

Incorporating this improved phase-type regression model within a joint modelling framework, as shall be detailed within this chapter, extends the scope of the Coxian phase-type distribution to scenarios where there exists some association between the survival or queueing process under investigation and a related longitudinal marker of interest. Currently, phase-type regression models have not been developed to allow for the inclusion of time-varying covariates, limiting their applicability to time-invariant scenarios, which is particularly detrimental within the medical field where it is often of interest to consider the effect of time-varying covariates on survival.

In turn, joint models can benefit from the distributional advantages of phase-type distributions, which can overcome some of the previously documented limitations of standard joint modelling approaches [19]. Namely, the ability of the Coxian to suitably represent any positive distribution to an arbitrary degree of accuracy overcomes the limitation of the more commonly employed parametric distributions, such as the exponential and Weibull, which have been noted by Gould [12], for example, to "restrict the range of baseline hazard functions that can be captured accurately". Indeed, this limitation has influenced the development of alternative spline-based [7, 161] and

piecewise constant [162] hazard representations, which can be more conceptually and computationally complex. Further, the fully parametric nature of the Coxian phase-type regression model means that the standard errors of the parameters within the joint likelihood are not underestimated, as they are when representing the survival process with a Cox PH model, due to its unspecified baseline [11]. It has additionally been highlighted within the literature that a fully parametric survival model is more convenient when interest lies in obtaining individualised predictions of survival outcome [12], which is attractive within the medical field due to the increasing popularity of 'personalised medicine'.

The Coxian phase-type distribution is also advantageous in comparison to alternative survival representations due to the additional information it can provide about the survival process, as previously discussed. That is to say, when modelling survival of a chronic condition, the latent phases which are uncovered during the model fitting can be considered to represent distinct stages of the disease under investigation. From these uncovered states, and the estimated rates of flow between them, invaluable insight can be obtained pertaining to how individuals behave before their failure time. For instance, within the E-step of the EM algorithm approach to fitting phase-type regression models, developed within Section 3.3, personalised approximations are made of the expected time each individual will spend in the underlying states of the system, providing information on how long individuals can expect to remain in the earlier stages of the disease, informing intervention strategies.

4.3 Two-stage Approach

Mirroring the first steps in the development of the standard joint modelling methodology, a two-stage approach was initially considered, similar to that of Tsiatis et al. [100] and Pawitan and Self [101], where the Coxian phase-type regression model was instead utilised to represent the survival process, in place of the standard Cox PH model. Consequently, within stage one, the longitudinal response is modelled to generate unbiased estimates of each individual's longitudinal response trajectory and, within stage two, features of this trajectory are incorporated within a survival model to quantify the association between the two processes.

Different parameterisations of the model can be formulated depending upon the features of the longitudinal trajectory which are incorporated within the survival model. The two most common parameterisations, (i) the random effects (RE) and (ii) the true longitudinal response (TLR), are detailed here. For illustration, the random effects parameterisation was further applied to data collected on individuals suffering from

chronic kidney disease, as detailed by Donnelly et al. [29].

4.3.1 Stage 1: Linear Mixed Effects Model

Within both parameterisations, the first stage of the model building procedure comprises of fitting a LME model to the repeated measures of the time-varying longitudinal response of interest, estimating population-level fixed effects parameters, as well as individual-specific deviations from these parameters, referred to as random effects. Most commonly, two random effects are incorporated within the model, representing deviations from the population-average intercept and slope, allowing individuals to have unique initial values of the longitudinal marker, and unique rates of change over time.

Fitting a LME model to the repeated observations of a periodically observed biomarker of interest makes the endogenous covariate fully observable, allowing a 'true' estimate to be calculated for each individual at their respective event times, free from contamination of unexplained variance.

The LME model, discussed in full within Section 2.2.3, is given by:

$$\begin{aligned} y_i(t) &= \boldsymbol{x}_i(t)\boldsymbol{\beta} + b_{i0} + b_{i1}t + \boldsymbol{\epsilon}_i(t) \\ &= y_i^*(t) + \epsilon_i(t) \end{aligned} \tag{4.1}$$

where b_{i0} and b_{i1} are the random effects and $y_i^*(t)$ is the unbiased estimate of the true longitudinal response at time t.

4.3.2 Stage 2: Coxian Phase-type Resgression Model

Employing the Coxian phase-type regression model to represent the survival process, replacing the standard Cox PH representation, allows additional information regarding the behaviour of the individuals before they experience their event of interest to be ascertained. That is, the phases of the Coxian phase-type distribution, and the rates of transition between them, provide insight into how quickly individuals will deteriorate through the underlying stages of the disease before experiencing their event of interest.

The two parameterisations of the Coxian phase-type regression model are detailed below, where, for simplicity in terms of the notation, the covariates are assumed to have a constant effect on each of the transitions through the Coxian. In practice, this assumption can be relaxed to make state-specific, direction-specific or transition-

specific inferences regarding the covariates' effects by alternatively utilising one of the formulations developed within Section 3.4.

4.3.2.1 Random Effects Parameterisation

Under the RE parameterisation, features of the individuals' longitudinal trajectories, represented by the latent random effects, are incorporated within the Coxian phase-type regression model, given by:

$$\log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - \widehat{b}_{i0}\alpha_1 - \widehat{b}_{i1}\tau_i\alpha_2 \Big\} E_{ij} + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} \Big(\log(q_{0jk}) - \boldsymbol{w}_i\boldsymbol{\gamma} - \widehat{b}_{i0}\alpha_1 - \widehat{b}_{i1}\tau_i\alpha_2 \Big) \Bigg\} \tag{4.2}$$

where $\sum_{j=1}^{n} B_{ij} \log(p_j)$, from the generalised log-likelihood of phase type distributions, given by Equation 3.21, is not included within the likelihood due to the assumption of the Coxian that all individuals begin the process within the first phase of the underlying system, meaning $B_{ij} \log(p_j) = 0$ for all i and j. The random effects parameterisation allows inferences to be made regarding the effect of deviating from the population average trajectory, in terms of both the intercept and slope, on the survival process, as was originally proposed by Tsiatis et al. [100].

4.3.2.2 True Longitudinal Response Parameterisation

Under the TLR parameterisation, the 'true' estimate of each individual's longitudinal response is incorporated as a covariate within the Coxian phase-type regression model, thus allowing its effect on the survival process to be evaluated, where additional baseline covariates can also be included, as shown:

$$\log L(\boldsymbol{\theta}_\tau; \tau_i) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - \widehat{y_i^*(\tau_i)}\alpha \Big\} E_{ij} + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} \Big(\log(q_{0jk}) - \boldsymbol{w}_i\boldsymbol{\gamma} - \widehat{y_i^*(\tau_i)}\alpha \Big) \Bigg\} \tag{4.3}$$

From the fitted model, it is therefore possible to make inferences on the effect of a one unit increase in the longitudinal response on the rates of flow through these latent stages of the survival process.

Some extensions within the literature have investigated the possibility of incorporating different representations of $y_i^*(\tau_i)$ within the survival model, which can similarly be extended to the Coxian phase-type regression model. For example, Crowther et al. [163] incorporated the first differential of the longitudinal response with respect to time, $y_i^*(\tau_i)'$, in order to make inferences regarding the effect of the rate of change of the longitudinal response on survival.

4.4 Joint Likelihood Approach

As previous research has shown, bias can be introduced to the estimated parameters from a two-stage joint modelling approach as no consideration is given to the potential impact of the survival process upon the longitudinal process, discussed fully within Section 2.4.1. Consequently, within this section, the parameters of both the longitudinal and survival processes are estimated simultaneously from a single joint likelihood, where the latent association between the two processes is represented by the random effects, as within the shared parameter formulation of joint models. Both the RE and TLR parameterisations are detailed within 4.4.1 and Sections 4.4.2 below, with the application of the TLR parameterisation to chronic kidney disease data illustrated within Chapter 5.

4.4.1 Random Effects Parameterisation

Under the RE parameterisation, the joint probability density of the longitudinal and survival processes, represented by a LME model and Coxian phase-type regression model respectively, is given by:

$$L(\boldsymbol{\theta}; \boldsymbol{y}_i, \tau_i) = \prod_{i=1}^{m} \int f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}_y) f(\tau_i|\boldsymbol{b}_i; \boldsymbol{\theta}_\tau) f(\boldsymbol{b}_i; \boldsymbol{\theta}_b) \mathrm{d}\boldsymbol{b}_i \tag{4.4}$$

where

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}_y) = \frac{1}{(2\pi\sigma^2)^{\frac{m_i}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\right\}, \tag{4.5}$$

$$f(\tau_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau) = \prod_{j=1}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\exp\left\{ - q_{0jk}\int_{F_{ij}}^{F_{i(j+1)}}\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du\right\}$$
$$\times\prod_{j=1}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\left(q_{0jk}\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}F_{i(j+1)}\alpha_2\Big\}\right)^{N_{ijk}} \tag{4.6}$$

$$f(\boldsymbol{b}_i;\boldsymbol{\theta}_b) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{D}|^{\frac{1}{2}}}\exp\left\{-\frac{\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i}{2}\right\}. \tag{4.7}$$

Note that, once again, the contribution to this density of the Coxian phase-type regression model, $f(\tau_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau)$, no longer includes the p_j or B_{ij} terms from the density of the standard phase-type regression model due to the assumption that all individuals begin the process within the first phase of the distribution, meaning $p_j^{B_{ij}} = 1$ for all i and j.

The corresponding log-likelihood of this probability density is given by:

$$\begin{aligned}\log L(\boldsymbol{\theta};\boldsymbol{y}_i,\tau_i) = \sum_{i=1}^{m}\Bigg\{ &- \frac{m_i}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\\ &- \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{D}|) - \frac{1}{2}\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i\\ &+ \sum_{j=1}^{n}\sum_{\substack{k=0\\k\neq j}}^{n} - q_{0jk}\int_{F_{ij}}^{F_{i(j+1)}}\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du\\ &+ \sum_{j=1}^{n}\sum_{\substack{k=0\\k\neq j}}^{n}N_{ijk}\Big(\log(q_{0jk}) - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}F_{i(j+1)}\alpha_2\Big)\Bigg\}\end{aligned} \tag{4.8}$$

where the EM algorithm is employed to maximise the likelihood and estimate the unknown parameters of the model, as detailed below.

4.4.1.1 E-Step

Within the E-step of the algorithm, the expected values of various functions of the random effects are approximated in a way similar to that outlined for standard joint models by Wulfsohn and Tsiatis [4] and Henderson et al. [9]. That is, on each iteration of the algorithm, the expected value of any function of the random effects, denoted $r(\boldsymbol{b}_i)$, is given by:

$$\begin{aligned}\mathbf{E}\Big[r(\boldsymbol{b}_i) \mid \boldsymbol{y}_i,\ \tau_i,\ \boldsymbol{b}_i;\boldsymbol{\theta}\Big] &= \int r(\boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{y}_i,\tau_i;\boldsymbol{\theta})\mathrm{d}\boldsymbol{b}_i \\ &= \frac{\int r(\boldsymbol{b}_i) f(\boldsymbol{y}_i,\tau_i,\boldsymbol{b}_i;\boldsymbol{\theta})\mathrm{d}\boldsymbol{b}_i}{f(\boldsymbol{y}_i,\tau_i;\boldsymbol{\theta})} \\ &= \frac{\int r(\boldsymbol{b}_i) f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_y) f(\tau_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau) f(\boldsymbol{b}_i;\boldsymbol{\theta}_b)\mathrm{d}\boldsymbol{b}_i}{\int f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_y) f(\tau_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau) f(\boldsymbol{b}_i;\boldsymbol{\theta}_b)\mathrm{d}\boldsymbol{b}_i}\end{aligned} \tag{4.9}$$

where $f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_y)$, $f(\tau_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau)$ and $f(\boldsymbol{b}_i;\boldsymbol{\theta}_b)$ are as defined in Equations 4.5, 4.6 and 4.7, and $r(\boldsymbol{b}_i)$ is replaced by the following required functions of the random effects:

$$r(\boldsymbol{b}_i) = \begin{cases} \boldsymbol{b}_i \\ \boldsymbol{b}_i'\boldsymbol{b}_i \\ \exp\Big\{ - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} \\ \boldsymbol{b}_i \exp\Big\{ - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} \\ \boldsymbol{b}_i'\boldsymbol{b}_i \exp\Big\{ - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\}. \end{cases} \tag{4.10}$$

The integrals with respect to the random effects can be approximated using the Gauss Hermite quadrature approach, or the pseudo-adaptive Gauss Hermite approach, detailed by Rizopoulos [17]. Within this research, the pseudo-adaptive approach was adopted.

The latent variables of the Coxian phase-type regression model are also calculated within the E-step, in a way that is similar to the standard phase-type regression model presented within Section 3.3.1, where additional derivatives of the log-likelihood with respect to α and γ are required for the M-step of the algorithm. The expected values of the latent Coxian phase-type regression variables are given by:

$$\mathbf{E}\Big[N_{ijk} \mid \boldsymbol{y}_i\ \tau_i,\ \boldsymbol{b}_i;\boldsymbol{\theta}\Big] = \frac{q_{0jk}c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.11}$$

$$\mathbf{E}\Big[N_{ij0} \mid \boldsymbol{y}_i\ \tau_i,\ \boldsymbol{b}_i;\boldsymbol{\theta}\Big] = \begin{cases} \dfrac{a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)q_{0j0}\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}\tau_i\alpha_2 \Big\}}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)}, & \text{if } \delta_i = 1 \\ 0, & \text{if } \delta_i = 0 \end{cases} \tag{4.12}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{c_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.13}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} b_{i1}u\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{h2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.14}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} (b_{i1}u)^2\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{h3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.15}$$

where $a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $c_{ijk}(\tau_i \mid \mathbf{T})$, $h2_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $h3_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$ and the elements of the vector $\mathbf{d_i}(\tau_i \mid \boldsymbol{\theta}_\tau)$, denoted $d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, are given by:

$$a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{p}\exp\left\{\mathbf{T}\int_0^{\tau_i}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du\right\}\mathbf{e_j} \tag{4.16}$$

$$\begin{aligned} d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) =& \mathbf{e_j}'\exp\left\{\mathbf{T}\int_0^{\tau_i}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du\right\} \\ &\times\left(\mathbf{t}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}\tau_i\alpha_2\Big\}\right)^{\delta_i} \end{aligned} \tag{4.17}$$

$$\begin{aligned} c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau) =& \int_0^{\tau_i}\mathbf{p}\exp\left\{\mathbf{T}\int_0^{u}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}s\alpha_2\Big\}ds\right\} \\ &\times\mathbf{e_j}\left(\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}\right)\mathbf{e_k}' \\ &\times\exp\left\{\mathbf{T}\int_u^{\tau_i}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}s\alpha_2\Big\}ds\right\} \\ &\times\left(\mathbf{t}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}\tau_i\alpha_2\Big\}\right)^{\delta_i}du \end{aligned} \tag{4.18}$$

$$\begin{aligned} h2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) =& \int_0^{\tau_i}\mathbf{p}\exp\left\{\mathbf{T}\int_0^{u}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}s\alpha_2\Big\}ds\right\} \\ &\times\mathbf{e_j}\left(b_{i1}u\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}\right)\mathbf{e_j}' \\ &\times\exp\left\{\mathbf{T}\int_u^{\tau_i}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}s\alpha_2\Big\}ds\right\} \end{aligned}$$

$$\times \left(\mathbf{t} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}\tau_i\alpha_2 \Big\} \right)^{\delta_i} du \tag{4.19}$$

$$h3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^u \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}s\alpha_2 \Big\} ds \right\}$$
$$\times \mathbf{e_j} \left((b_{i1}u)^2 \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} \right) \mathbf{e_j}'$$
$$\times \exp\left\{ \mathbf{T} \int_u^{\tau_i} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}s\alpha_2 \Big\} ds \right\}$$
$$\times \left(\mathbf{t} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}\tau_i\alpha_2 \Big\} \right)^{\delta_i} du. \tag{4.20}$$

where the integrals with respect to time are numerically approximated using the Gauss-Kronrod approach [123].

4.4.1.2 M-Step

By differentiating the log-likelihood with respect to the parameters of interest and solving, updated estimates of the fixed effects parameters, $\boldsymbol{\beta}$, the variance of the random effects parameter, $\mathbf{D}$, and the variance of the residual errors parameter, σ^2, are given by:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \mathbf{X_i}'\mathbf{X_i} \right)^{-1} \left(\sum_{i=1}^{m} (\mathbf{y_i} - \mathbf{Z_i}\widehat{\boldsymbol{b}}_i)'\mathbf{X_i} \right) \tag{4.21}$$

$$\widehat{\mathbf{D}} = \frac{1}{m} \sum_{i=1}^{m} \left[\widehat{\boldsymbol{b}}_i \widehat{\boldsymbol{b}}_i{}' + \mathrm{Var}\big(\widehat{\boldsymbol{b}}_i\big) \right] \tag{4.22}$$

$$\widehat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^{m} (\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})'(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta} - 2\mathbf{Z_i}\widehat{\boldsymbol{b}}_\mathbf{i}) + \mathrm{trace}\Big(\mathbf{Z_i}'\mathbf{Z_i}\mathrm{Var}\big(\widehat{\boldsymbol{b}}_i\big) \Big) + \widehat{\boldsymbol{b}}_i{}'\mathbf{Z_i}'\mathbf{Z_i}\widehat{\boldsymbol{b}}_i. \tag{4.23}$$

where the features of the random effects are approximated within the E-step.

Similarly, updated estimates of the baseline transition intensity parameters, q_{0jk}, are given by:

$$\begin{aligned}\widehat{q}_{0jk} &= \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m} \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du} \\ \widehat{q}_{0jj} &= -\sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk}.\end{aligned} \tag{4.24}$$

where $\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{\boldsymbol{w}_i\boldsymbol{\gamma} + b_{i0}\alpha_1 + b_{i1}u\alpha_2\Big\}du$ is approximated within the E-step, given by Equation 4.13.

Closed form estimates of the survival parameters, α_1, α_2 and $\boldsymbol{\gamma}$, cannot be obtained, meaning a one-step Newton Raphson is implemented to updated these parameters, which, on the $(l+1)^{th}$ iteration are given by:

$$\widehat{\boldsymbol{\gamma}}^{(l+1)} = \widehat{\boldsymbol{\gamma}}^{(l)} - H\Big(\widehat{\boldsymbol{\gamma}}^{(l)}\Big)^{-1} S\Big(\widehat{\boldsymbol{\gamma}}^{(l)}\Big) \tag{4.25}$$

$$\widehat{\alpha}_1^{(l+1)} = \widehat{\alpha}_1^{(l)} - H\Big(\widehat{\alpha}_1^{(l)}\Big)^{-1} S\Big(\widehat{\alpha}_1^{(l)}\Big) \tag{4.26}$$

$$\widehat{\alpha}_2^{(l+1)} = \widehat{\alpha}_2^{(l)} - H\Big(\widehat{\alpha}_2^{(l)}\Big)^{-1} S\Big(\widehat{\alpha}_2^{(l)}\Big) \tag{4.27}$$

where $S(\cdot)$ and $H(\cdot)$ denotes the score vector and Hessian matrix, respectively.

The score vector and Hessian matrix of the baseline survival parameters, denoted $S(\boldsymbol{\gamma})$ and $H(\boldsymbol{\gamma})$, are given by:

$$S(\boldsymbol{\gamma}) = \sum_{i=1}^{m}\left\{ \sum_{j=1}^{n}\sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk}\boldsymbol{w}_i \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du - \sum_{j=1}^{n}\sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk}\boldsymbol{w}_i \right\} \tag{4.28}$$

$$H(\boldsymbol{\gamma}) = \sum_{i=1}^{m}\left\{ \sum_{j=1}^{n}\sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk}\boldsymbol{w}_i\boldsymbol{w}_i' \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2\Big\}du \right\} \tag{4.29}$$

where $\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du$ is approximated within the E-step, given by Equation 4.13.

The score and Hessians of the α_1 and α_2 parameters, representing the effects of the features of the longitudinal response parameters on the survival process, are given by:

$$S(\alpha_1) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk} b_{i0} \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} b_{i0} \Bigg\} \tag{4.30}$$

$$H(\alpha_1) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} {b_{i0}}^2 \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du \Bigg\} \tag{4.31}$$

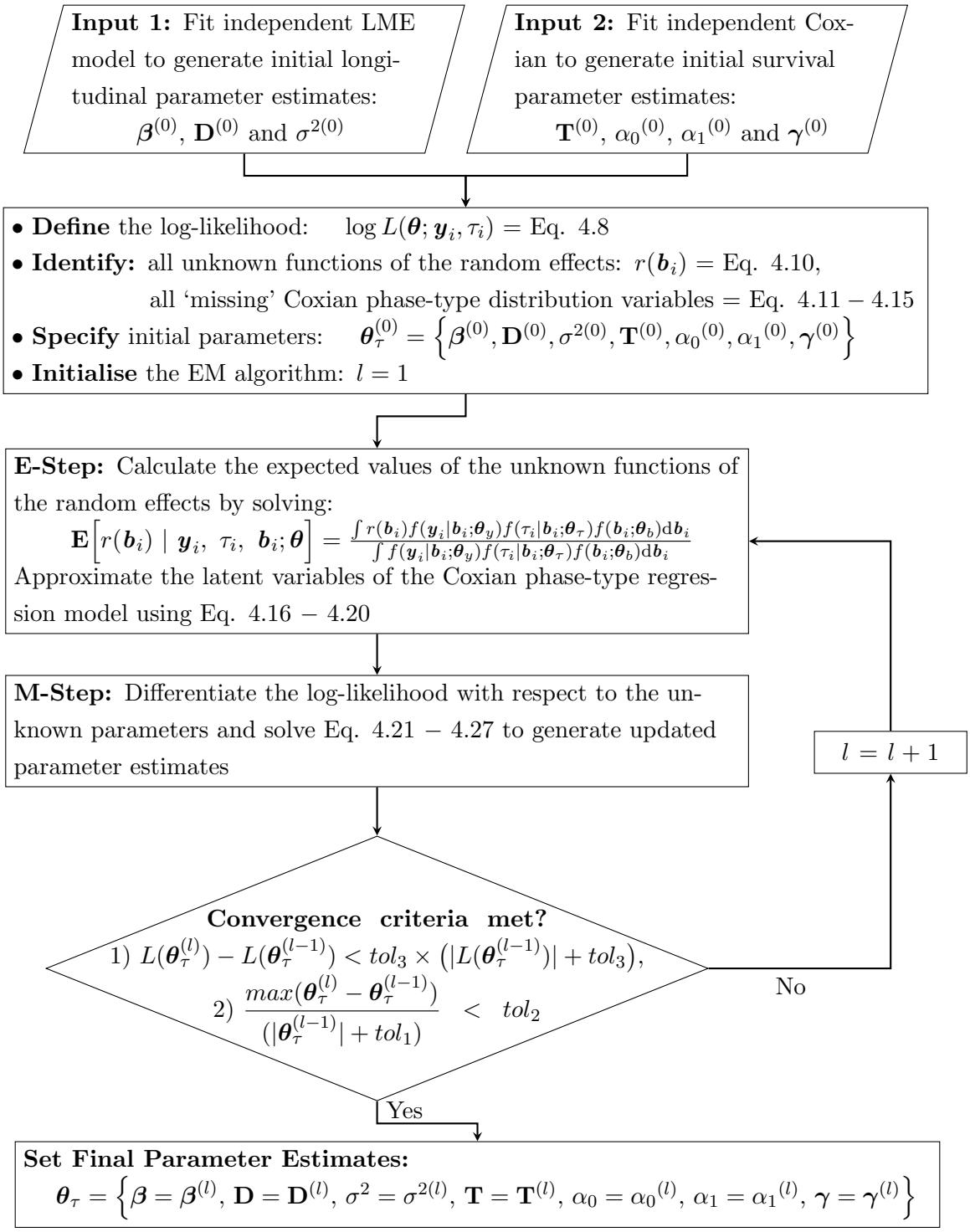
and

$$S(\alpha_2) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} b_{i1}u \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} b_{i1} F_{i(j+1)} \Bigg\} \tag{4.32}$$

$$H(\alpha_2) = \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} \big(b_{i1}u\big)^2 \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du \Bigg\} \tag{4.33}$$

where the integrals within these equations, $\int_{F_{ij}}^{F_{i(j+1)}} b_{i1}u \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du$ and $\int_{F_{ij}}^{F_{i(j+1)}} \big(b_{i1}u\big)^2 \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - b_{i0}\alpha_1 - b_{i1}u\alpha_2 \Big\} du$, are again approximated within the E-step, given by Equations 4.14 and 4.15.

4.4.1.3 Summary



4.4.2 True Longitudinal Response Parameterisation

Under the TLR parameterisation, the joint probability density of the longitudinal and survival processes, represented by a LME model and Coxian phase-type regression model respectively, is again given by Equation 4.4, where $f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_y)$ and $f(\boldsymbol{b}_i;\boldsymbol{\theta}_b)$ remain unchanged from the RE parameterisation, given by Equations 4.5 and 4.7 respectively. However, under this parameterisation, the contribution of the survival process to this density is given by:

$$f(\tau_i|\boldsymbol{b}_i;\boldsymbol{\theta}_\tau) = \prod_{j=1}^{n}\prod_{\substack{k=0\\k\neq j}}^{n} \exp\left\{ -q_{0jk}\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\}du\right\} \times \prod_{j=1}^{n}\prod_{\substack{k=0\\k\neq j}}^{n}\left(q_{0jk}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(F_{i(j+1)})\alpha\Big\}\right)^{N_{ijk}}, \tag{4.34}$$

where $y_i^*(\cdot) = \boldsymbol{x}_i(\cdot)\boldsymbol{\beta} + \boldsymbol{z}_i(\cdot)\boldsymbol{b}_i$. The corresponding log-likelihood of this complete probability density function is thus given by:

$$\begin{aligned}\log L(\boldsymbol{\theta};\boldsymbol{y}_i,\tau_i) = \sum_{i=1}^{m}\Bigg\{ &-\frac{m_i}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\\ &-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{D}|) - \frac{1}{2}\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i\\ &+\sum_{j=1}^{n}\sum_{\substack{k=0\\k\neq j}}^{n} -q_{0jk}\int_{F_{ij}}^{F_{i(j+1)}}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\}du\\ &+\sum_{j=1}^{n}\sum_{\substack{k=0\\k\neq j}}^{n} N_{ijk}\Big(\log(q_{0jk}) - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(F_{i(j+1)})\alpha\Big)\Bigg\}\end{aligned} \tag{4.35}$$

and the EM algorithm approach is again utilised to maximise this likelihood, as described within the next two sections below.

4.4.2.1 E-Step

Within the E-step, similarly to the RE parameterisation, the expected values of various functions of the random effects, $r(\boldsymbol{b}_i)$, are approximated utilising the method of

Wulfsohn and Tsiatis [4], given by Equation 4.9, where the required functions of the random effects, $r(\boldsymbol{b}_i)$, are:

$$r(\boldsymbol{b}_i) = \begin{cases} \boldsymbol{b}_i \\ \boldsymbol{b}_i'\boldsymbol{b}_i \\ \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} \\ y_i^*(u)\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} \\ y_i^*(u)^2\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\}. \end{cases} \tag{4.36}$$

The latent variables of the Coxian phase-type regression model are also calculated within the E-step, in the say way as within the RE parameterisation discussed in Section 4.4.1, where additional derivatives of the log-likelihood with respect to β are required for the M-step of the algorithm. The expected values of the latent Coxian phase-type regression variables are given by:

$$\mathbf{E}\Big[N_{ijk} \mid \boldsymbol{y}_i\ \tau_i,\ \boldsymbol{b}_i; \boldsymbol{\theta}\Big] = \frac{q_{0jk}c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.37}$$

$$\mathbf{E}\Big[N_{ij0} \mid \boldsymbol{y}_i\ \tau_i,\ \boldsymbol{b}_i; \boldsymbol{\theta}\Big] = \begin{cases} \dfrac{a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)q_{0j0}\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha \Big\}}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)}, & \text{if } \delta_i = 1 \\ 0, & \text{if } \delta_i = 0 \end{cases} \tag{4.38}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{c_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.39}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} y_i^*(u)\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{h2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.40}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} y_i^*(u)^2\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{h3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.41}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} \boldsymbol{x}_i(u)\alpha\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{h4_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.42}$$

$$\mathbf{E}\left[\int_{F_{ij}}^{F_{i(j+1)}} \boldsymbol{x}_i(u)'\boldsymbol{x}_i(u)\alpha^2\exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \mid \tau_i,\ \boldsymbol{w}_i\right] = \frac{h5_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \tag{4.43}$$

where $a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $h2_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $h3_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $h4_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$, $h5_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau)$

and the elements of the vector $\boldsymbol{d}_i(\tau_i \mid \boldsymbol{\theta}_\tau)$, denoted $d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)$, are given by:

$$a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \right\} \mathbf{e_j} \tag{4.44}$$

$$d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau) = \mathbf{e_j}' \exp\left\{ \mathbf{T} \int_0^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \right\} \left(\mathbf{t} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha \Big\} \right)^{\delta_i} \tag{4.45}$$

$$\begin{aligned} c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau) = & \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{u} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\} \\ & \times \mathbf{e_j} \left(\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} \right) \mathbf{e_k}' \\ & \times \exp\left\{ \mathbf{T} \int_u^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\} \left(\mathbf{t} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha \Big\} \right)^{\delta_i} du \end{aligned} \tag{4.46}$$

$$\begin{aligned} h2_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = & \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{u} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\} \\ & \times \mathbf{e_j} \left(y_i^*(u) \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} \right) \mathbf{e_j}' \\ & \times \exp\left\{ \mathbf{T} \int_u^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\} \left(\mathbf{t} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha \Big\} \right)^{\delta_i} du \end{aligned} \tag{4.47}$$

$$\begin{aligned} h3_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = & \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{u} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\} \\ & \times \mathbf{e_j} \left(y_i^*(u)^2 \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} \right) \mathbf{e_j}' \\ & \times \exp\left\{ \mathbf{T} \int_u^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\} \left(\mathbf{t} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha \Big\} \right)^{\delta_i} du \end{aligned} \tag{4.48}$$

$$h4_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p} \exp\left\{ \mathbf{T} \int_0^{u} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha \Big\} ds \right\}$$

$$\times \mathbf{e_j}\left(\boldsymbol{x}_i(u)\alpha \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\}\right)\mathbf{e_j}' \times \exp\left\{\mathbf{T}\int_u^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha\Big\}ds\right\}\left(\mathbf{t}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha\Big\}\right)^{\delta_i} du \tag{4.49}$$

$$h5_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau) = \int_0^{\tau_i} \mathbf{p}\exp\left\{\mathbf{T}\int_0^{u} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha\Big\}ds\right\} \times \mathbf{e_j}\left(\boldsymbol{x}_i(u)'\boldsymbol{x}_i(u)\alpha^2 \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\}\right)\mathbf{e_j}' \times \exp\left\{\mathbf{T}\int_u^{\tau_i} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(s)\alpha\Big\}ds\right\}\left(\mathbf{t}\exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(\tau_i)\alpha\Big\}\right)^{\delta_i} du. \tag{4.50}$$

4.4.2.2 M-Step

Updated estimates of the variance of the random effects parameter, $\mathbf{D}$, and the variance of the residual errors parameter, σ^2, can be calculated similarly to the RE parameterisation, utilising Equations 4.22 and 4.23. Similarly, updated estimates for the baseline transition intensity parameters, q_{0jk}, are give by:

$$\begin{aligned}\widehat{q}_{0jk} &= \frac{\sum_{i=1}^{m} N_{ijk}}{\sum_{i=1}^{m}\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\}du} \\ \widehat{q}_{0jj} &= -\sum_{\substack{k=0\\k\neq j}}^{n} q_{0jk},\end{aligned} \tag{4.51}$$

where $\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ -\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\}du$ is approximated within the E-step, given by Equation 4.39.

As the longitudinal fixed effects parameters, $\boldsymbol{\beta}$, are now contained within the survival portion of the probability density, closed form estimates can no longer be obtained. Instead, estimates of $\boldsymbol{\beta}$, along with the survival parameters α and $\boldsymbol{\gamma}$, are obtained through a one-step Newton Raphson update:

$$\widehat{\boldsymbol{\beta}}^{(l+1)} = \widehat{\boldsymbol{\beta}}^{(l)} - H\Big(\widehat{\boldsymbol{\beta}}^{(l)}\Big)^{-1} S\Big(\widehat{\boldsymbol{\beta}}^{(l)}\Big) \tag{4.52}$$

$$\widehat{\boldsymbol{\gamma}}^{(l+1)} = \widehat{\boldsymbol{\gamma}}^{(l)} - H\Big(\widehat{\boldsymbol{\gamma}}^{(l)}\Big)^{-1} S\Big(\widehat{\boldsymbol{\gamma}}^{(l)}\Big) \tag{4.53}$$

$$\widehat{\alpha}^{(l+1)} = \widehat{\alpha}^{(l)} - H\Big(\widehat{\alpha}^{(l)}\Big)^{-1} S\Big(\widehat{\alpha}^{(l)}\Big) \tag{4.54}$$

The score and Hessian of the fixed effects parameters, denoted $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ respectively, are given by:

$$\begin{aligned} S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \Bigg\{ & \frac{1}{\sigma^2} \mathbf{X_i}'(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\widehat{\boldsymbol{b}}_i) \\ & + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} \boldsymbol{x}_i(u)\alpha \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \\ & - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} \boldsymbol{x}_i(F_{i(j+1)})\alpha \Bigg\} \end{aligned} \tag{4.55}$$

$$\begin{aligned} H(\boldsymbol{\beta}) = \sum_{i=1}^{m} \Bigg\{ & - \frac{1}{\sigma^2} (\mathbf{X_i}'\mathbf{X_i}) \\ & + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} \boldsymbol{x}_i(u)'\boldsymbol{x}_i(u)\alpha^2 \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \Bigg\} \end{aligned} \tag{4.56}$$

where the integrals within the score and Hessian, $\int_{F_{ij}}^{F_{i(j+1)}} \boldsymbol{x}_i(u)\alpha \exp\Big\{-\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\} du$ and $\int_{F_{ij}}^{F_{i(j+1)}} \boldsymbol{x}_i(u)'\boldsymbol{x}_i(u)\alpha^2 \exp\Big\{-\boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha\Big\} du$, are approximated within the E-step of the algorithm and are given by Equations 4.42 and 4.43.

The score vector and Hessian matrix of the baseline survival parameters, denoted $S(\boldsymbol{\gamma})$ and $H(\boldsymbol{\gamma})$, are given by:

$$S(\boldsymbol{\gamma}) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk} \boldsymbol{w}_i \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du + \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} \boldsymbol{w}_i \right\} \tag{4.57}$$

$$H(\boldsymbol{\gamma}) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \boldsymbol{w}_i \boldsymbol{w}_i' \int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \right\} \tag{4.58}$$

where the integral $\int_{F_{ij}}^{F_{i(j+1)}} \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du$ is approximated within the E-step, given by Equation 4.39.

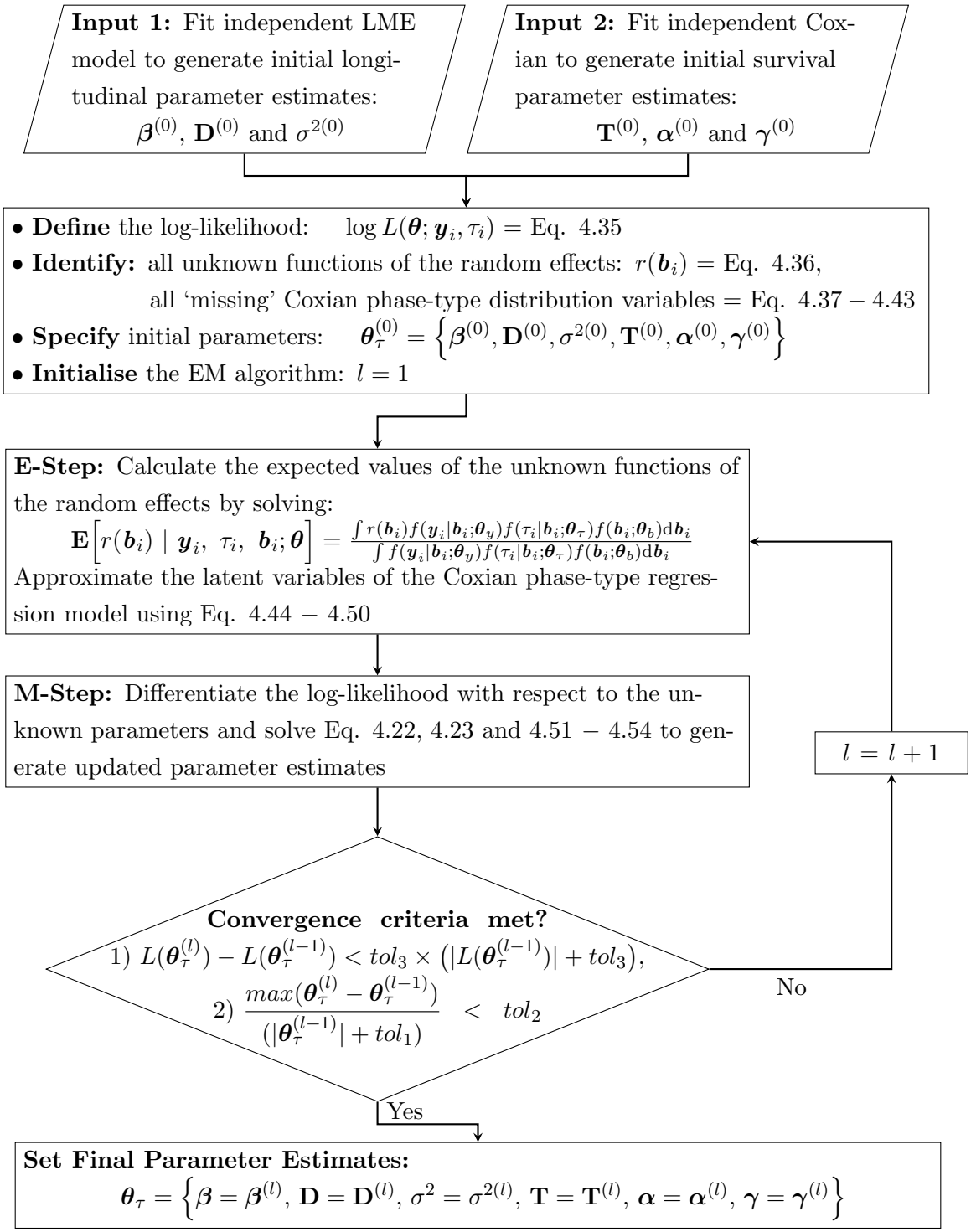
The score vector and Hessian matrix of the association parameter of the true longitudinal response and survival outcome, denoted $S(\alpha)$ and $H(\alpha)$, are given by:

$$S(\alpha) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} y_i^*(u) \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \right.$$
$$\left. - \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} N_{ijk} y_i^*\big(F_{i(j+1)}\big) \right\} \tag{4.59}$$

$$H(\alpha) = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} \sum_{\substack{k=0 \\ k\neq j}}^{n} -q_{0jk} \int_{F_{ij}}^{F_{i(j+1)}} y_i^*(u)^2 \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du \right\} \tag{4.60}$$

where $\int_{F_{ij}}^{F_{i(j+1)}} y_i^*(u) \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du$ and $\int_{F_{ij}}^{F_{i(j+1)}} y_i^*(u)^2 \exp\Big\{ - \boldsymbol{w}_i\boldsymbol{\gamma} - y_i^*(u)\alpha \Big\} du$ are estimated within the E-step and are given by Equations 4.40 and 4.41.

4.4.2.3 Summary



4.4.3 Simulation Study Three

A simulation study was implemented to validate the TLR parameterisation of the new joint likelihood approach which utilises the Coxian phase-type regression model to represent the survival process, as developed within Section 4.4.2. Within this study, two scenarios are considered. In the first scenario, the survival process is simulated from a two-phase Coxian, whilst in the second scenario the survival process is simulated from a three-phase Coxian, with 20% censoring in each. In doing so, the ability of the model to handle an increasing number of underlying phases (and thus an increasing number of unknown parameters) is illustrated. Censoring times were simulated for each individual from a similar survival distribution, appropriately adjusted to ensure the correct proportion of censoring, where the earliest event time (death or censor) was taken for each individual. In order to demonstrate the potential error that occurs within the estimates of the survival parameters when the underlying distribution is misspecified, joint models which assume an exponential and Weibull distribution, the standard within the JM package, are also fitted to the simulated datasets.

For both the two- and three-phase simulations, 100 datasets were produced, and each of the joint model formulations under investigation (i.e. the Coxian, exponential, Weibull AFT joint models) were fitted to the data. The average parameter estimates of these 100 fits, along with their standard errors and empirical confidence intervals (CI), are then compared to the true simulated values of the parameters. Convergence was considered to be achieved if either of the following conditions, previously employed within the JM package, was satisfied [14]:

i $L(\boldsymbol{\theta}_\tau^{(l)}) - L(\boldsymbol{\theta}_\tau^{(l-1)}) < tol_3 \times \big(|L(\boldsymbol{\theta}_\tau^{(l-1)})| + tol_3\big)$, or

ii $\dfrac{max(\boldsymbol{\theta}_\tau^{(l)} - \boldsymbol{\theta}_\tau^{(l-1)})}{(|\boldsymbol{\theta}_\tau^{(l-1)}| + tol_1)} < tol_2,$

where $tol_1 = 1 \times 10^{-3}$ and $tol_2 = 1 \times 10^{-4}$, and $tol_3 = 1 \times 10^{-8}$, and a maximum of 1000 iterations of the algorithm were permitted.

4.4.3.1 Two-Phase Simulation

For the two-phase simulation study, 100 datasets were generated, each comprised of 400 individuals with an average of 13.19 repeated measures per individual, randomly distributed between time zero and the individuals' event times. For each individual, the observed longitudinal response of interest at time t_{ij} was calculated by:

$$y_i(t_{ij}) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + x_{i2}\beta_2 + x_{i3}\beta_3 + \epsilon_{ij} \tag{4.61}$$

where:

t_{ij} is the j^{th} observation time for individual i, where the observation times were initially generated uniformly between 0 and the 99^{th} percentile of the distribution, and subsequently truncated based upon the simulated event time,

x_{i2} is a continuous covariate generated from a uniform distribution bounded between -3 and 3: $x_{i2} \sim unif(-3, 3)$,

x_{i3} is a binary covariate generated from a discrete uniform distribution bounded between 0 and 1: $x_{i3} \sim unif\{0, 1\}$,

β_0, β_1, β_2 and β_3 are the population level regression parameters corresponding to the fixed effects, given by $\boldsymbol{\beta} = \begin{pmatrix} -10.00 & 2.10 & -0.40 & 0.50 \end{pmatrix}'$,

b_{i0} and b_{i1} are the random effects, representing the individuals' deviations from the population average intercepts and slopes, respectively, generated from a multivariate normal distribution;

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix}, \begin{pmatrix} 0.80 & 0.15 \\ 0.15 & 0.30 \end{pmatrix} \right],$$

ϵ_{ij} represents the residual error on the j^{th} observation from the i^{th} individual, generated from a normal distribution: $\epsilon_{ij} \sim \mathcal{N}(0.00, 0.30)$.

For the survival process, the individuals' event times were simulated according to a two-phase Coxian distribution with baseline transition parameters $q_{010} = 0.0$, $q_{012} = 0.1$ and $q_{020} = 0.3$, influenced by two covariates:

$$q_{ijk}(t) = q_{0jk} \exp\big\{ - y_i(t)^* \alpha \big\} \exp\big\{ - w_i \gamma \big\},$$

where:

$y_i(t)^*$ is the true value of the longitudinal response of interest at time t, given by:

$$y_i^*(t) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t + x_{i2}\beta_2 + x_{i3}\beta_3 \tag{4.62}$$

with corresponding association parameter $\alpha = 0.60$, and,

w_i, a continuous baseline survival covariate generated from a uniform distribution, $w_i \sim unif(-3, 3)$, with corresponding regression parameter $\gamma = -0.25$.

The mean parameter estimates from the fits of the three joint models which assume an underlying two-phase Coxian, exponential and Weibull distribution, along with their standard errors, CIs and bias, are given within Tables 4.1 and 4.2.

Table 4.1: Parameter estimates from the two-phase Coxian joint model fitted to the simulated datasets.

Par	**Sim**	**Two-phase Coxian**			
		Est.	**Std Err.**	**95% CI**	**Bias (%)**
β_0	-10.000	-9.987	0.006	-10.065, -9.894	+0.013 (0.130)
β_1	2.100	2.102	0.002	2.069, 2.140	+0.002 (0.095)
β_2	-0.400	-0.404	0.002	-0.441, -0.363	−0.004 (1.000)
β_3	0.500	0.490	0.005	0.378, 0.584	−0.010 (2.000)
σ^2	0.300	0.301	0.001	0.289, 0.315	+0.001 (0.333)
D(1,1)	0.800	0.797	0.006	0.659, 0.918	−0.003 (0.375)
D(1,2)	0.300	0.298	0.002	0.244, 0.344	−0.002 (0.667)
D(2,2)	0.150	0.149	0.001	0.127, 0.169	−0.001 (0.667)
α	0.600	0.621	0.004	0.550, 0.679	+0.021 (3.500)
γ	-0.250	-0.263	0.005	-0.365, -0.188	−0.013 (5.200)

Par: Parameter, Sim: True Simulated Value, Est.: Mean Estimated Value,
Std. Err.: Standard Error, CI: Empirical Confidence Intervals
Bias = Mean Estimated Value − True Simulated Value
Red: Difference between Mean Estimated Value and True Simulated Value when Est.<Sim
Blue: Difference between Mean Estimated Value and True Simulated Value when Est.>Sim

From these results it can be observed that, for each of the three models, the longitudinal and variance parameters were estimated with no significant bias. However, the estimates of the survival parameters can be seen to be influenced by the choice of survival distribution, where the exponential and, to a lesser extent, Weibull models produced biased estimates which overestimated the covariate effects. In comparison, the two-phase Coxian showed no significant bias in the estimates of the survival parameters.

The estimated baseline survival probability densities of the three parametric models are plotted within Figure 4.1, alongside that of the simulated density, illustrating that only the two-phase Coxian successfully uncovered the true shape of the distribution. This highlights how the exponential and Weibull distributions struggle when the true survival distribution is complex, as is likely to be the case with real world data.

Table 4.2: Parameter estimates from the two-phase Coxian joint model fitted to the simulated datasets.

Par	Sim	Exponential				Weibull			
		Est.	**Std Err.**	**95% CI**	**Bias (%)**	**Est.**	**Std Err.**	**95% CI**	**Bias (%)**
β_0	-10.000	-9.990	0.006	-10.081, -9.888	+0.010 (0.100)	-10.001	0.006	-10.082, -9.889	−0.001 (0.010)
β_1	2.100	2.100	0.002	2.069, 2.137	0.000 (0.000)	2.098	0.002	2.069, 2.137	−0.002 (0.095)
β_2	-0.400	-0.405	0.002	-0.439, -0.365	−0.005 (1.250)	-0.402	0.002	-0.439, -0.365	−0.002 (0.500)
β_3	0.500	0.491	0.008	0.370, 0.586	−0.009 (1.800)	0.507	0.008	0.370, 0.535	+0.007 (1.400)
σ^2	0.300	0.301	0.001	0.290, 0.315	+0.001 (0.333)	0.301	0.001	0.290, 0.315	+0.001 (0.333)
D(1,1)	0.800	0.799	0.007	0.692, 0.918	−0.001 (0.125)	0.798	0.006	0.688, 0.914	−0.002 (0.250)
D(1,2)	0.300	0.298	0.002	0.256, 0.341	−0.002 (0.667)	0.298	0.003	0.255, 0.341	−0.002 (0.667)
D(2,2)	0.150	0.148	0.001	0.128, 0.168	−0.002 (1.333)	0.148	0.001	0.129, 0.169	−0.002 (1.333)
α	0.600	0.840	0.003	0.782, 0.912	+0.240 (40.00)	0.647	0.024	0.414, 1.159	+0.047 (7.83)
γ	-0.250	-0.353	0.009	-0.454, -0.252	−0.103 (41.20)	-0.276	0.012	-0.409, -0.152	−0.026 (10.40)

Par: Parameter, Sim: True Simulated Value, Est.: Mean Estimated Value, Std. Err.: Standard Error, CI: Empirical Confidence Intervals
Bias = Mean Estimated Value − True Simulated Value
Red: Difference between Mean Estimated Value and True Simulated Value when Est.<Sim
Blue: Difference between Mean Estimated Value and True Simulated Value when Est.>Sim

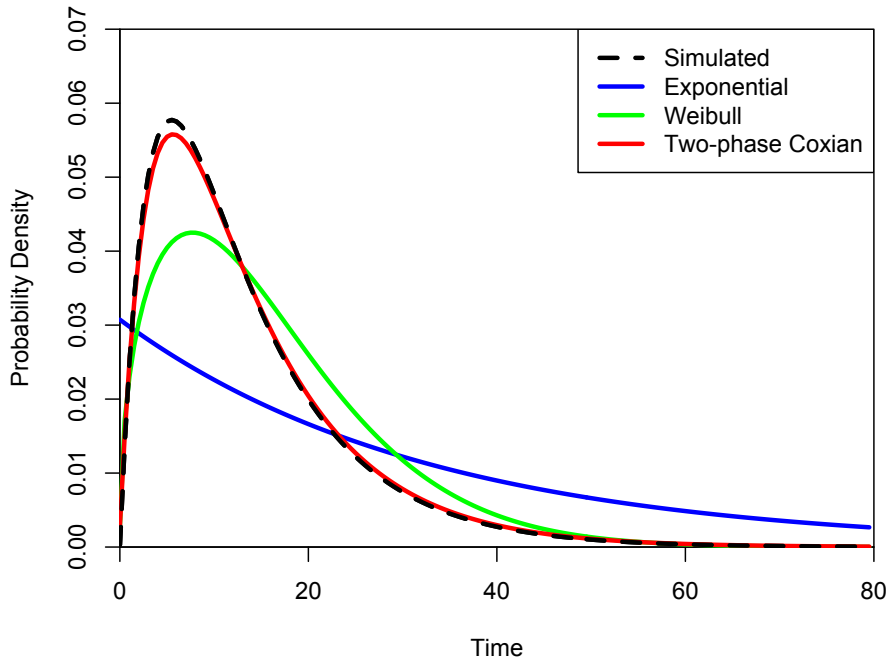


Figure 4.1: Plot showing the estimated baseline probability density functions when the survival process is represented by an exponential, Weibull, and two-phase Coxian distribution.

4.4.3.2 Three-Phase Simulation

For the three-phase simulation study, 100 datasets were generated, comprised of 450 individuals with an average of 5.09 repeated measures per individual. The observed longitudinal response of interest at time t_{ij}, denoted $y_i(t_{ij})$, was again defined by four fixed effects: (i) an intercept, (ii) the time of observation, (iii) a continuous covariate generated from a uniform distribution, $x_{i2} \sim unif(-3, 3)$, and (iv) a binary covariate also from a uniform distribution, $x_{i3} \sim unif\{0, 1\}$. The corresponding fixed effects regression parameters are given by: $\boldsymbol{\beta} = \begin{pmatrix} 2.00 & 3.50 & 0.20 & -0.60 \end{pmatrix}$.

Similarly to the two phase simulation, random individual-level variation was introduced to the longitudinal response through two multivariate normally distributed random effects, b_{i0} and b_{i1}, defined by:

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix}, \begin{pmatrix} 0.60 & -0.20 \\ -0.20 & 0.10 \end{pmatrix} \right],$$

and observation level residual errors, ϵ_{ij}, were also introduced, generated from a normal distribution: $\epsilon_{ij} \sim \mathcal{N}(0.00, 0.40)$.

For the survival process, the individuals' event times were simulated according to a three-phase Coxian distribution with baseline transition parameters $q_{010} = 0.00$, $q_{020} = 0.40$, $q_{030} = 0.02$, $q_{012} = 0.14$ and $q_{023} = 0.40$. As before, the individual-specific transition intensities are influenced by the longitudinal process, with association parameter $\alpha = 0.20$, and a continuous baseline covariate with regression parameter $\gamma = 0.10$.

The mean parameter estimates from the fits of the joint model assuming a three-phase Coxian, an exponential, and a Weibull distribution, along with their standard errors, CIs and bias, are given within Tables 4.3 and 4.4. For the three-phase Coxian, it can be observed that all parameters were estimated with minimal bias. For the exponential and Weibull joint models however, it can be seen that whilst these formulations successfully estimated the longitudinal parameters, they once again did not appropriately uncover the true simulated values of the survival parameters γ and α, as was the case within the two-phase simulation.

Table 4.3: Parameter estimates from the three-phase Coxian joint model fitted to the simulated datasets.

Par	**Sim**	**Two-phase Coxian**			
		Est.	**Std Err.**	**95% CI**	**Bias (%)**
β_0	2.000	1.995	0.005	1.912, 2.097	−0.005 (0.250)
β_1	3.500	3.503	0.002	3.473, 3.546	+0.003 (0.086)
β_2	0.200	0.201	0.002	0.169, 0.232	+0.001 (0.500)
β_3	-0.600	-0.612	0.006	-0.707, -0.492	−0.012 (2.000)
σ^2	0.400	0.400	0.001	0.372, 0.421	0.000 (0.000)
D(1,1)	0.600	0.595	0.006	0.489, 0.759	−0.005 (0.833)
D(1,2)	0.100	0.103	0.001	0.080, 0.131	+0.003 (3.000)
D(2,2)	-0.200	-0.203	0.003	-0.277, -0.160	−0.003 (1.500)
α	0.200	0.180	0.006	0.106, 0.272	−0.020 (10.000)
γ	0.100	0.091	0.005	0.012, 0.178	−0.009 (9.000)

Par: Parameter, Sim: True Simulated Value, Est.: Mean Estimated Value
Std. Err.: Standard Error, CI: Empirical Confidence Intervals
Bias = Mean Estimated Value − True Simulated Value
Red: Difference between Mean Estimated Value and True Simulated Value when Est.<Sim
Blue: Difference between Mean Estimated Value and True Simulated Value when Est.>Sim

Interestingly, in this case, it was the exponential rather than the Weibull distribution which provided estimates which are closer to the true values of the parameters.

Table 4.4: Parameter estimates from the standard exponential and Weibull AFT joint models fitted to the simulated datasets.

Par	Sim	Exponential				Weibull			
		Est.	**Std Err.**	**95% CI**	**Bias (%)**	**Est.**	**Std Err.**	**95% CI**	**Bias (%)**
β_0	2.000	2.000	0.005	1.902, 2.096	0.000 (0.000)	2.007	0.005	1.907, 2.104	+0.007 (0.350)
β_1	3.500	3.499	0.002	3.467, 3.539	−0.001 (0.029)	3.492	0.002	3.456, 3.533	−0.008 (0.229)
β_2	0.200	0.203	0.002	0.174, 0.242	+0.003 (1.500)	0.203	0.002	0.173, 0.242	+0.003 (1.500)
β_3	-0.600	-0.599	0.005	-0.708, -0.488	+0.001 (0.167)	-0.598	0.005	-0.706, -0.488	+0.002 (0.333)
σ^2	0.400	0.400	0.001	0.374, 0.424	0.000 (0.000)	0.400	0.001	0.374, 0.424	0.000 (0.000)
D(1,1)	0.600	0.604	0.006	0.389, 0.698	+0.004 (0.667)	0.604	0.006	0.494, 0.716	+0.004 (0.667)
D(1,2)	0.100	0.103	0.001	0.070, 0.133	+0.003 (3.000)	0.103	0.001	0.078, 0.131	+0.003 (3.000)
D(2,2)	-0.200	-0.205	0.003	-0.249, -0.122	−0.005 (2.500)	-0.205	0.003	-0.264, -0.157	−0.005 (2.500)
α	0.200	0.154	0.001	0.136, 0.173	−0.046 (23.00)	0.030	0.002	-0.006, 0.061	−0.170 (85.00)
γ	0.100	0.077	0.004	-0.002, 0.167	−0.023 (23.00)	0.040	0.002	-0.001, 0.087	−0.060 (60.00)

Par: Parameter, Sim: True Simulated Value, Est.: Mean Estimated Value, Std. Err.: Standard Error, CI: Empirical Confidence Intervals
Bias = Mean Estimated Value − True Simulated Value
Red: Difference between Mean Estimated Value and True Simulated Value when Est.<Sim
Blue: Difference between Mean Estimated Value and True Simulated Value when Est.>Sim

This is not unsurprising as the estimated baseline density of the exponential distribution is closer to that of the true simulated density than the Weibull distribution, as can be seen within Figure 4.2. However, as was the case for the two-phase simulation, only the Coxian fit uncovered the true shape of the distribution.

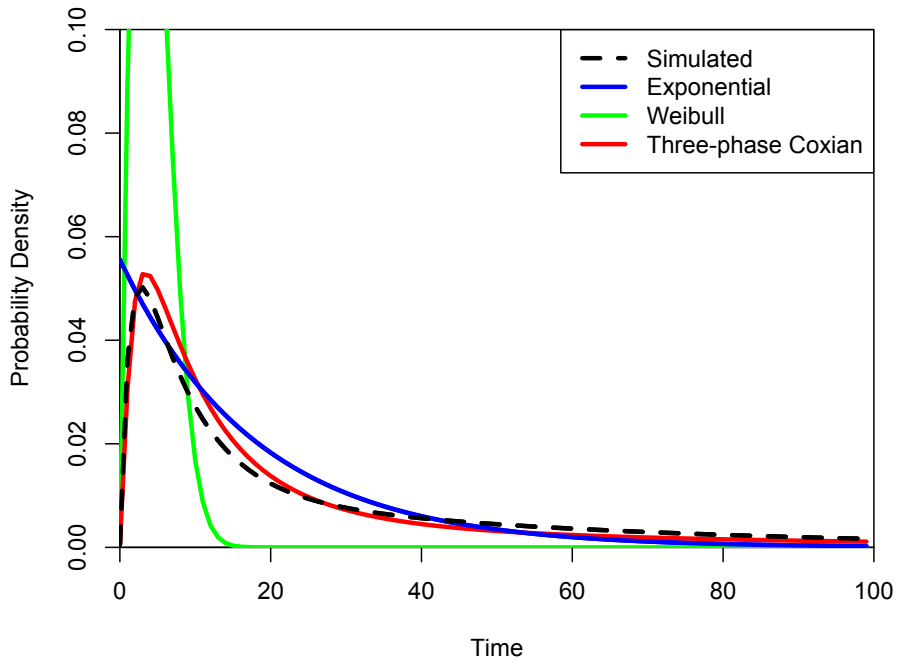


Figure 4.2: Estimated baseline probability density functions when survival is represented by an exponential, Weibull, and three-phase Coxian distribution.

4.5 Summary

Within this chapter, a new joint modelling framework was developed within which the longitudinal process is represented by a LME model and the survival process by a Coxian phase-type regression model. Initially, a two stage approach was explored, motivated by the early approaches to incorporate unbiased estimates of a time-varying, endogenous biomarker within a survival model. Subsequently, the joint likelihood approach to simultaneously estimate the parameters of both submodels, under both the random effects (RE) and true longitudinal response (TLR) parameterisations, was detailed in full.

A simulation study was then performed utilising the TLR parameterisation, illustrating the suitability of the Coxian to represent the survival process within such a

joint modelling framework, where the simulated parameters were successfully uncovered by the new model. Further, the simulation study also demonstrates the potential bias which can be introduced to the estimates of the survival parameters when the underlying distribution is misspecified, as joint models which assumed an exponential and Weibull distribution were shown to not properly uncover the true covariate effects.

The novel joint model developed within this chapter contributes significantly to both the areas of joint modelling and phase-type distributions in a number of ways:

i Employing the Coxian phase-type regression model to represent survival within a joint likelihood enables joint models to uncover latent stages of the survival process, meaning more insight can be obtained regarding the process under investigation. As mentioned previously, within a disease modelling context, such insight can be utilised to predict rates of deterioration of patients through the uncovered stages of the disease, informing treatment interventions and providing predictions on quality of life.

ii As the Coxian phase type distribution can represent any positive distribution to an arbitrary degree of accuracy, it is not subject to the misspecification issues associated with the exponential and Weibull distributions, which are limited by the distributional shapes which they can represent. As illustrated within the simulation study, miss-specifying the distribution can cause bias within the estimates of the survival parameters, which the Coxian was shown to overcome.

iii Incorporating the Coxian phase-type distribution within a joint framework significantly extends the scope of the Coxian phase-type regression model, allowing the incorporation of time-varying, endogenous covariates as predictors for the first time within the literature. Consequently, phase-type distributions now constitute a more appropriate survival model, applicable to cases where time-varying covariates are of interest.

Chapter 5

Application to Chronic Kidney Disease Patients

5.1 Overview

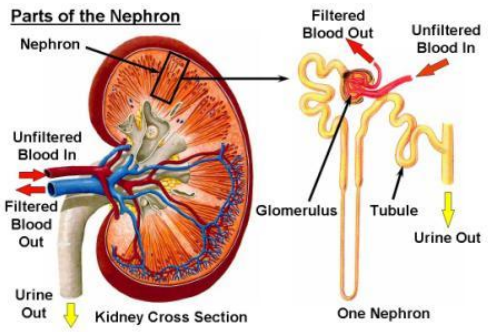
Within this chapter, the newly developed approach to incorporate the Coxian phase-type regression model within a joint modelling framework, developed within Chapter 4, is applied to a dataset collected on individuals suffering from chronic kidney disease (CKD), illustrating its applicability within disease modelling. Section 5.2 first provides some biological background on CKD and outlines the targets of this investigation, with the dataset subsequently introduced within Section 5.3 and some preliminary data analysis presented. Within Section 5.4, the process of fitting the model to the data is discussed and the results are presented. The new methodology is also compared to standard joint modelling approaches currently found within the literature, highlighting its advantages and ability to overcome limitations of the previous techniques.

5.2 Biological Background

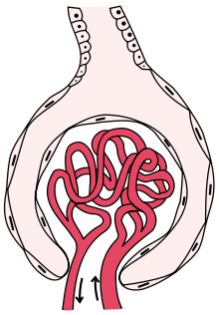
5.2.1 Chronic Kidney Disease

The kidneys are responsible for the cleaning and regulation of the body's blood supply, where their primary function is to filter waste products from the blood and convert the waste into urine which is expelled from the body [164]. They are also responsible for maintaining blood pressure, producing erythropoietin (EPO); a chemical which stimulates the production of red blood cells, and for maintaining the correct level of chemicals within the blood [165]. The filtration of the body's blood supply takes place inside tubular structures located within the kidneys called nephrons. At the beginning of each nephron there is a network of capillaries known as the glomerulus; blood travels through these capillaries at high pressure and plasma is filtered out through the capillary walls into a surrounding structure called Bowman's capsule. The rate at which at which blood is filtered through the glomeruli in this way is known as the glomerular filtration rate (GFR), and it is this which serves as one of the primary markers for overall renal function [166].

Kidney disease arises when, for some underlying reason, the kidneys have a reduced ability to carry out their functions, and it can occur in both an acute state or as a chronic condition. Acute kidney disease is a short term illness whereby kidney function is lost quickly, typically as a result of injury. However, when the underlying cause is treated, kidney function typically returns to normal with no lasting impacts. In contrast, chronic kidney disease is a long term, irreversible condition whereby an individual's kidney health gradually deteriorates over time [169].



Cross-section of a kidney highlighting a single nephron [167]



The structure of a single glomerulus [168]

Figure 5.1: Structure of the kidney, a single nephron and Bowman's capsule

In the United Kingdom, it has been estimated by a National Health Service (NHS) Kidney Care Report [170] that there are currently 1.8 million people diagnosed with CKD in England, costing the NHS approximately £1.45 billion. This is more than double the cost from 2002/2003, making CKD a prevailing challenge for healthcare providers [35]. The report also highlights that 95% of this cost can be attributed to secondary care of CKD patients, particularly renal replacement therapies, such as dialysis treatment. Consequently, it is mutually beneficial to both patients and healthcare providers to more accurately model the behaviour of CKD so as to provide treatment interventions with greater accuracy in a more cost-efficient manner.

Clinicians consider there to be five stages of CKD, and individuals are categorised into one of these five stages based on their estimated glomerular filtration rate (eGFR) as shown in Table 5.1.

Table 5.1: Stages of CKD

Stage	eGFR	Description
1	≥ 90	Normal
2	60-89	Mild reduction
3a	45-59	Mild - moderate reduction
3b	30-44	Moderate - severe reduction
4	15-29	Severe reduction
5	<15	Kidney failure

The eGFR is a measure of how well the kidneys filter waste from the blood, which can be difficult to assess accurately in routine clinical practice. One approach to estimate the GFR is to add an exogenous marker, such as iothalmate [171], to the

blood and to measure how much remains after the blood has passed through the kidneys. However, estimating GFR in this way can be complex and expensive, so instead it is more common for an endogenous marker, i.e. a waste product naturally produced by the body, to be used. The most common maker utilised is creatinine, a breakdown by-product of muscle metabolism found in the blood. Routine blood tests can give a measure of the level of creatinine found in the blood which is then used to calculate an estimate of the eGFR using the MDRD equation, which is given by [172]:

$$\begin{aligned} \text{eGFR} = 175 \times \text{standardised } S_{\text{Cr}}^{-1.154} \times \text{Age}^{-0.203} \\ \times (1.210 \text{ if Black}) \times (0.742 \text{ if Female}) \end{aligned} \tag{5.1}$$

where S_{Cr} is serum creatinine.

The individuals considered within this research are in end-stage renal failure and are receiving regular haemodialysis (HD) treatment. HD is a common intervention performed on end-stage CKD patients, which typically involves diverting the patients blood out of their body through a machine which filters waste products and excess fluid from the blood, artificially fulfilling the function of the ailing kidneys.

5.2.2 Anaemia

It is commonly observed that individuals suffering from CKD are also prone to suffer from anaemia, a condition characterised by a reduction in the oxygen-carrying capacity of the blood [173], typically as a result of fewer red blood cells or a decreased volume of haemoglobin (Hb) [174]. Anaemia is common within CKD patients due to the disease hindering the kidneys' normal secretion of EPO, the hormone responsible for regulating the production of red blood cells.

As CKD progresses, the production of EPO typically decreases and anaemia becomes increasingly prevalent within individuals suffering from severe kidney disease [175, 176]. As such, within this research, Hb is investigated as a potential marker for CKD, where interest lies in modelling the association between the dynamic nature of individuals Hb levels, and their survival outcome.

5.3 The Dataset

5.3.1 Introduction

The dataset analysed within this research was provided by the Northern Ireland Renal Information Service and it contains 27,113 observations collected on 1,340 individuals within Northern Ireland who are suffering from CKD and receiving haemodialysis (HD) treatment. The data was collected from eight treatment centres across Northern Ireland between April 2002 and December 2011, with a maximum survival followup until September 2012. Repeated measures were collected on the individual's Hb levels, where the average number of observations per individual is 20.2, with a maximum of 133.

Within this analysis, the individuals' repeatedly observed Hb levels are regressed upon various baseline biomarkers, and the association between the dynamic nature of Hb and survival time is investigated. An overview of the key variables observed within the dataset is discussed below.

Observation Time

Observation time was recorded in months from when patients commenced HD treatment. The maximum observed time spent on HD was 138.4 months (11.5 years), with a mean of 29.6 months (2.5 years) and median 22.4 months (1.9 years).

Age

The mean age of patients when they began HD treatment was 66.9 years old, with a median of 70, suggesting a slight negative skew which can be observed in the plotted distribution within Figure 5.2, suggesting CKD to be more common amongst older people. This is consistent with renal literature which highlights the disease as being particularly problematic within an ageing population [38, 177].

The youngest person observed within the dataset is 19 years old, and the oldest is 97, with the majority of patients (54.9%) aged between 65 and 85 years old. Within the analysis, age is measured per 10 years, where the baseline was set to 7, corresponding to a 70 year-old individual.

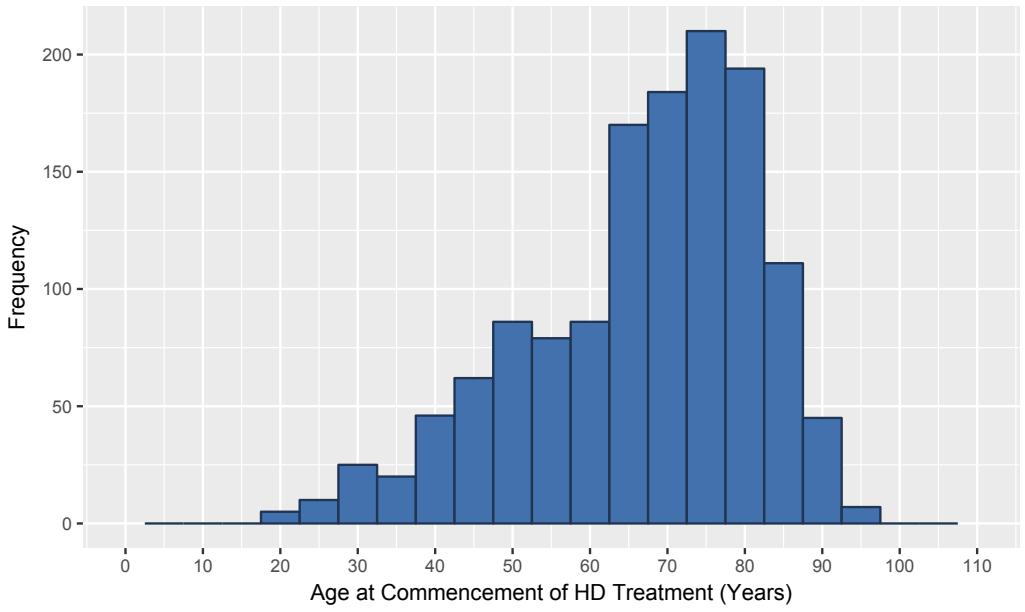


Figure 5.2: Histogram showing the distribution of age at commencement of HD treatment.

Gender

The dataset contained 523 (39%) females, who are defined as the baseline, and 817 (61%) males. This is again consistent with renal literature where it has been suggested that, despite a higher proportion of females being diagnosed with CKD, end-stage renal failure is more common in men [178].

Ethnicity

Whilst ten ethnic categories were recorded within the dataset (including 'Other'), there exists a major imbalance amongst the classes as 1278 patients (95.4%) were recorded as 'White'. There were 42 individuals (3.1%) for whom ethnicity was not recorded, leaving the remaining 20 individuals (1.5%) spread across nine categories, as shown in Table 5.2.

In order to overcome this imbalance, within all subsequent analysis only two ethnic groups were considered; 'White' and 'Other', where 'Other' is redefined to include all non-white ethnicities and is set as the baseline within the analysis.

Table 5.2: Frequency table showing the number of individuals within each ethnicity category

Ethnicity	Frequency	Percent
White	1278	95.4
Asian	3	0.22
Bangladesh	1	0.07
Black African	1	0.07
Chinese	3	0.22
Indian	1	0.07
Irish	7	0.52
Other Asian	1	0.07
Pakistani	2	0.15
Other	1	0.07
Missing	42	3.13

MCHC

Mean corpuscular haemoglobin concentration (MCHC) is a measure of the average concentration of haemoglobin in a given volume of packed red blood cells, measured in grams per deciliter (g/dL) [173]. The healthy adult range is between $33 - 36$ g/dL, where lower values can indicate anaemia due to iron deficiency. The observed mean (and median) within the data is 32.4 g/dL, indicating that the CKD patients have MCHC values slightly below the healthy range, as would be expected. The minimum observed value was 28.3 g/dL and the maximum was 38.2 g/dL, giving a small overall range of 9.9 g/dL, and interquartile range of 1.6 g/dL. Within the analysis, the baseline was redefined as 32 g/dL.

MCV

Mean corpuscular volume (MCV) is a measure of the average red blood cell size, recorded in femtoliters (fL), where the healthy adult range is between $80 - 96$ fL [173]. The mean (and median) observed within the data is 95 fL which lies close to the upper boundary of the acceptable levels. This is consistent with literature which has suggested that normocytic anaemia due to renal and chronic diseases has less impact on MCV levels, which typically remain within the normal range, in comparison to anaemia due to other causes [175]. The minimum observed value was 77.9 fL and the maximum was 128.8 fL, giving an overall range of 50.9 fL, with a small interquartile range of 8.2 fL. Within the analysis, MCV was measured per 10 fL, where the baseline was set to 9, corresponding to 90 fL.

Creatinine

Creatinine is a waste product produced by the body's muscle metabolism which is filtered from the body by the kidneys. It is a common endogenous marker used to calculate the eGFR of the kidneys and it is measured in micromoles per litre (μM/L), where the healthy adult range is between $60 - 110$ μM/L. Within the dataset the observed mean was 625.7 μM/L (median 605 μM/L) with a large overall range from 73 μM/L to 1567 μM/L, and a large interquartile range of 288.5. Due to its large scale, creatinine was analysed per 100 μM/L, where the baseline was defined as 6, corresponding to 600 μM/L.

Ferritin

Ferritin is a protein responsible for storing and releasing iron. Measured in nanograms per milliliter (ng/ml), the healthy range is $500 - 2000$ ng/ml [179]. Within the NI dataset, the overall mean ferritin level is slightly below the healthy range at 496.27 ng/ml, with a median of 403.5 ng/ml. The minimum observed level within the dataset was 15 ng/ml and the largest was 18030 ng/ml, giving an overall range of 18015 ng/ml, with and interquartile range of 403.35. Again due to its large scale, ferritin was analysed per 100 ng/ml, and the baseline was set to 4, corresponding to 400 ng/ml.

Urea

Urea is a waste product produced by the body and removed from the blood by the kidneys. Whilst higher levels than average can indicate reduced kidney function, urea is generally considered a poor marker for CKD as it is strongly influenced by diet and hydration levels at the time of observation [180]. Measured in millimoles per litre (mmol/L), the healthy range is between $2.5 - 7.1$ mmol/L, where the observed mean within the data is 18.0 mmol/L, which is set as the baseline, with a median of 17.7 mmol/L. The minimum observed value was 3.7 mmol/L and the maximum was 46.8 mmol/L, giving an overall range of 43.1 mmol/L, with a small interquartile range of 7.0 mmol/L.

Iron and EPO Treatments

The individuals observed within the NI Renal dataset received drug treatment interventions to improve their iron and EPO levels. The prescribed treatments, along with the number of individuals on each treatment, are listed within Table 5.4.

Summary Table

Tables 5.3 and 5.4 provide summaries of the key continuous and categorical variables observed within the NI Renal dataset, respectively.

Table 5.3: Summary of the continuous variables observed within the CKD dataset under investigation

Variable	Mean	Median	Min	Max
Age	66.9	70.0	19.0	97.0
MCMC	32.4	32.4	28.3	38.2
MCV	95.0	95.0	77.9	128.8
Creatinine	625.7	605.0	73.0	1567.0
Ferritin	496.3	403.5	15.0	18030.0
Urea	18.0	17.7	3.7	46.8

Table 5.4: Summary of the categorical variables observed within the CKD dataset under investigation

Variable	Level	Frequency	Percent
Gender	Male	817	61.0
	Female*	523	39.0
Ethnicity	White	1278	95.4
	Other*	62	4.6
Iron Treatment	Iron Hydroxide	269	20.1
	Venofer	755	56.3
	No Iron*	316	23.6
EPO Treatment	Aranesp	680	50.8
	EpoetinBeta	579	43.2
	Other*	81	6.0

* Indicates baseline.

5.3.2 Preliminary Data Analysis

Longitudinal Response of Interest: Haemoglobin

Haemoglobin (Hb) is an iron-rich protein found within red blood cells to which oxygen binds so as to be transported around the body. As discussed previously, anaemia is a common CKD comorbidity, where individuals can suffer from a reduced volume of both red blood cells and/or Hb. As such, Hb can be considered a marker for the individuals' anaemic condition, which varies in a way that reflects the underlying health of their kidneys. Measured in grams per deciliter (g/dL), deviations outside of the healthy range of $10.5 - 12.5$ g/dL can have a negative effect on the survival of

CKD patients [181].

The caterpillar plot in Figure 5.3 shows the mean and interquartile ranges of the (ordered) individuals' Hb levels. It can be observed that there exists considerable variation amongst the individuals' mean Hb levels, suggesting that the repeated measures are clustered by individual, as would be anticipated. This is further validated by calculating the intraclass-correlation coefficient (ICC), given by:

$$\text{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2} = \frac{0.51}{2.01} = 0.25 \tag{5.2}$$

where σ_b^2 represents the individual level variance and σ_ϵ^2 is the observation level variance, where these values are obtained by fitting a null LME model with a random intercept term to the data. From this ICC it can be inferred that, of the total variation which exists amongst the individuals' Hb levels, 25% can be attributed to variation amongst the individuals.

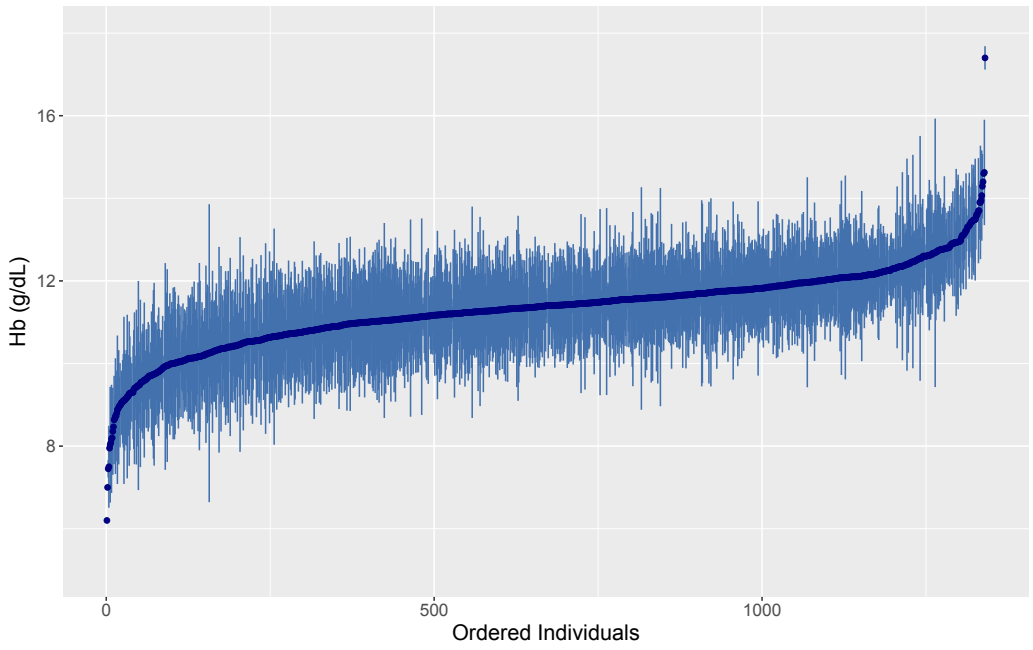


Figure 5.3: A caterpillar plot showing a large amount of individual-level variation amongst Hb levels observed within the sample.

This variation can be further observed by examining the trajectories of the individuals' Hb levels, stratified by future event outcome and plotted within Figures 5.4 and 5.5.

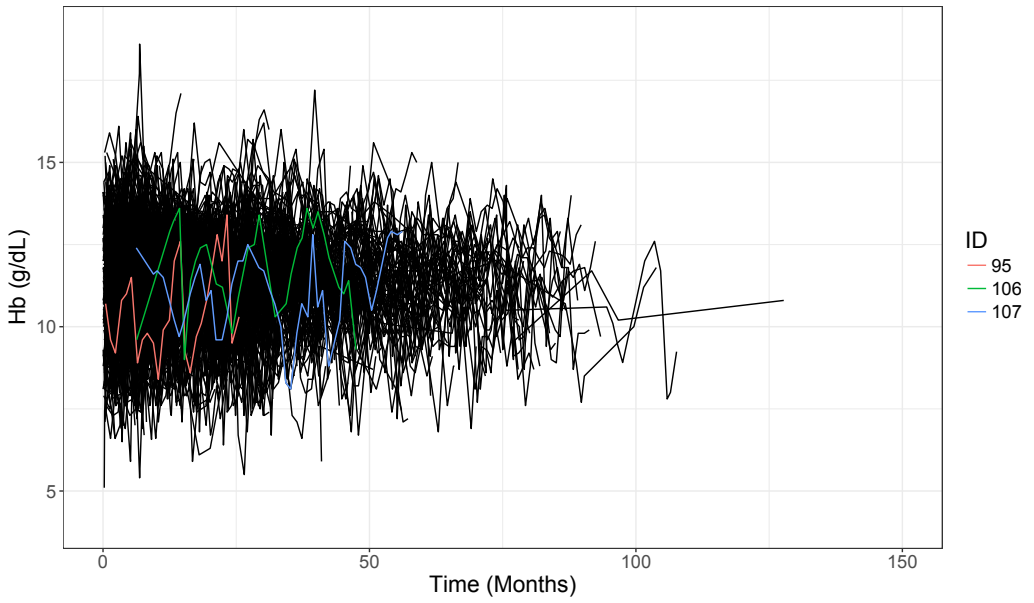


Figure 5.4: Spaghetti plot showing the Hb trajectories for those individuals who die during the observation period.

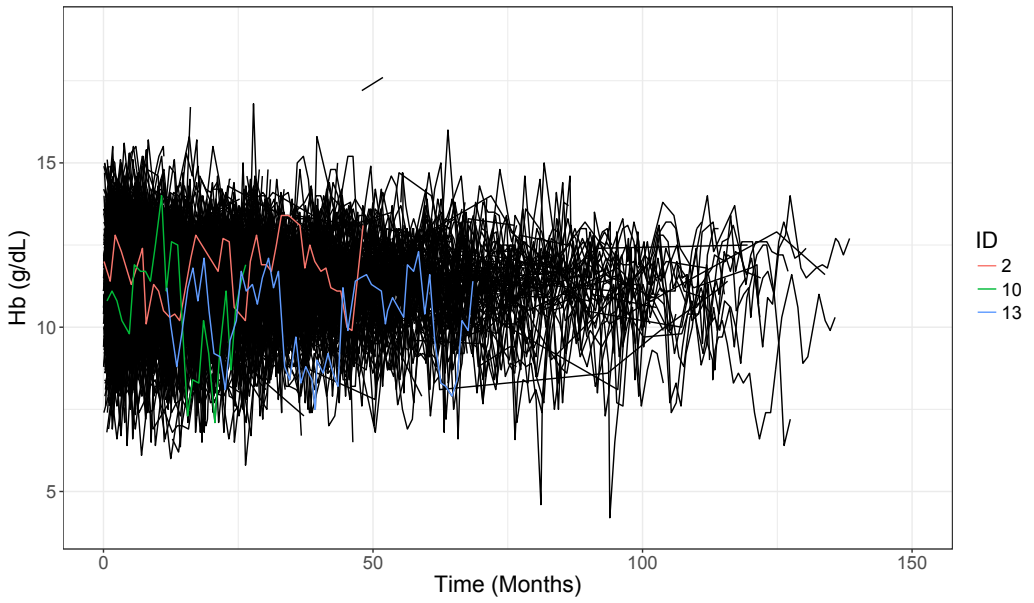


Figure 5.5: Spaghetti plot showing the Hb trajectories for those individuals who are censored within the dataset.

From these plots, it can be observed that not only does there exist a large amount of individual-level variation, but also a high level of variation amongst the repeated measures per individual, evidenced by the large fluctuations observed within the highlighted individuals' trajectories. Whilst it is difficult, within these plots at least, to identify much difference between the two strata, it can be seen that the censored individuals were observed for a longer period of time, with 298 observations made after 100 months on the censored individuals, compared to only 9 observations on those who die.

5.3.2.1 Survival Outcome: Time to death since commencement of HD

Within the dataset 585 (43.66%) individuals died during the observation period and 755 (56.34%) were censored. The distribution of the observed death times is shown in Figure 5.6 and a KM plot of the overall survival probability is shown in Figure 5.7.

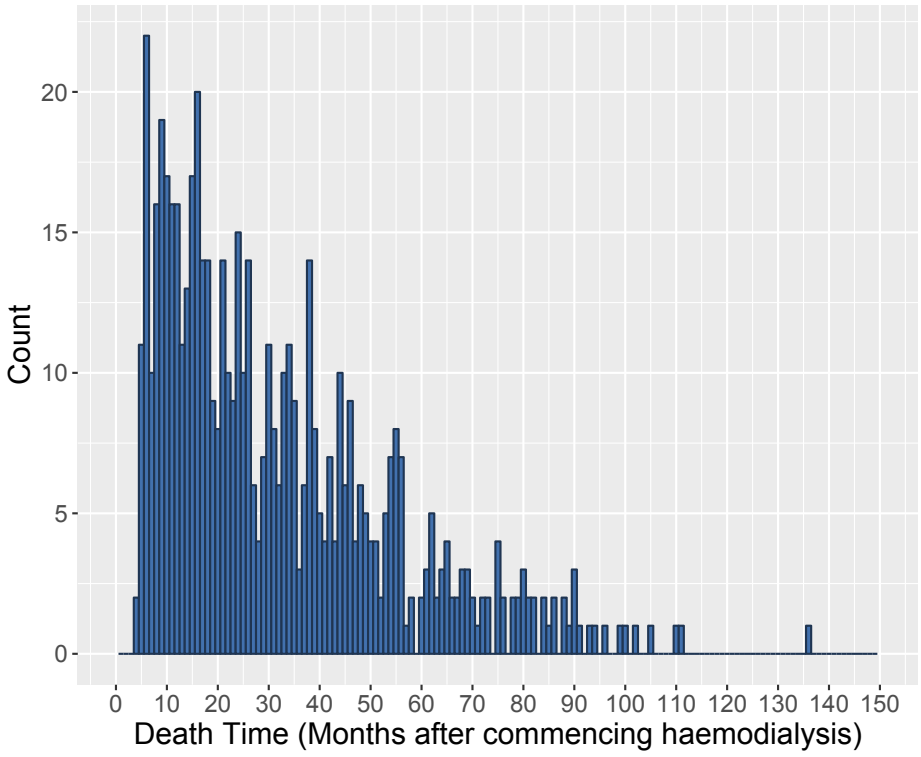


Figure 5.6: Histogram showing the distribution of the observed event times within the dataset.

Further, based upon advice from a clinician, the individuals' initial Hb levels were

discretised into 4 categories, as shown:

- Extremely Low Hb: < 9 g/dL
- Low Hb: $9 - 10.5$ g/dL
- Desired Hb: $10.5 - 12.5$ g/dL
- High Hb: > 12.5 g/dL

and KM survival plots for these four categories are given in Figure 5.8.

From Figure 5.8, it can be observed that individuals with low levels of Hb when commencing HD have worse survival probabilities compared to those with desired or high Hb levels. Fitting a Cox PH model utilising the individuals' initial Hb levels as the sole predictor of survival further suggest a possible association, where a one unit increase in Hb results in a hazard ratio (HR) of 0.891 (p-value: < 0.001), i.e. the hazard of death decreases as Hb increases, which is consistent with previous renal literature [182]. Of course, this test is prone to various sources of bias and error; it only utilises the individuals' initial Hb levels, which are prone to measurement error, and does not incorporate any of the subsequent longitudinal measurements, which are possibly informative of survival. Additionally, the effect of other significant variables is not controlled for within the model. However, the model serves to give an idea of the possible association which exists between Hb and survival, motivating subsequent analysis utilising more appropriate techniques, explored in full within Section 5.4.

5.4 Application of Statistical Models to NI Dataset

5.4.1 Independent Analysis of Longitudinal and Survival Data

Preliminarily, independent longitudinal and survival models were fitted to the NI data, providing an indication of the variables which significantly influence each of the processes, as well as supplying suitable initial values for the parameters within the joint model. It can also be of interest to compare the parameter estimates from the independent models with those from the joint model to identify which parameters are most prone to the bias which results from the naive assumptions of the independent models.

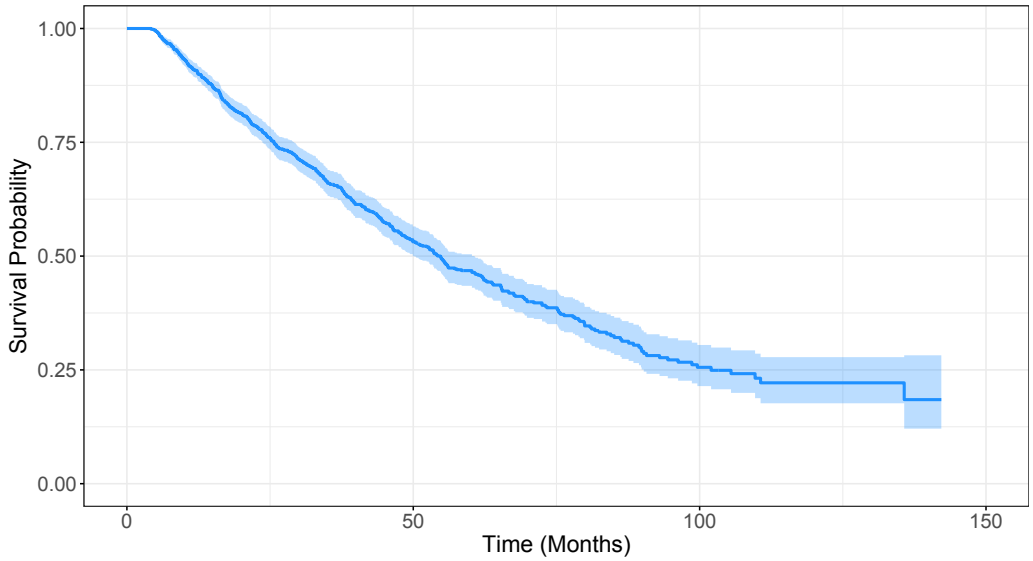


Figure 5.7: A KM plot showing the overall survival probability of the sample.

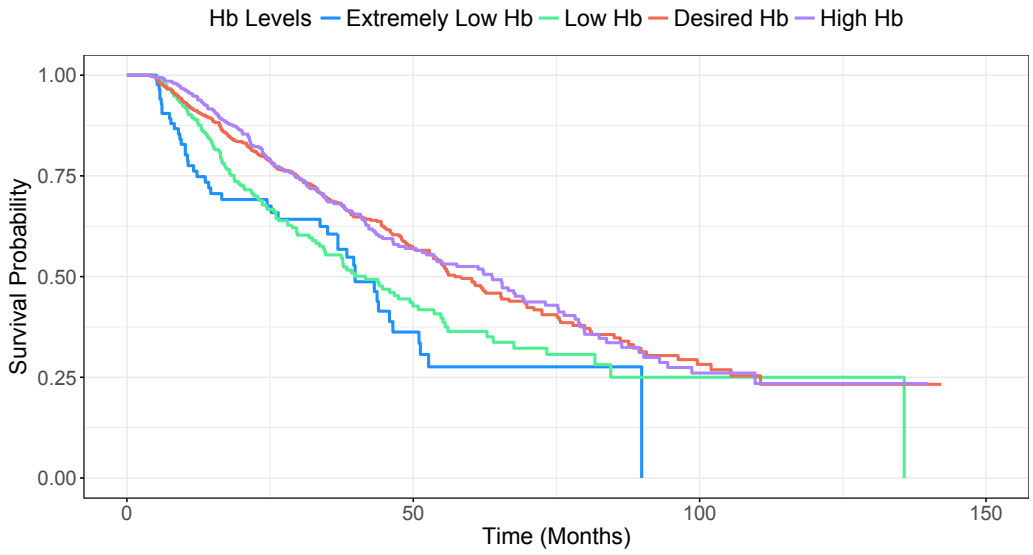


Figure 5.8: A KM plot showing the different survival probabilities for each of the four Hb categories.

5.4.1.1 LME Model for the Longitudinal Response

As discussed previously, LME models are employed when there exists significant clustering amongst the repeated measures of some longitudinal response of interest, violating the independent assumption of ordinary linear regression. Whilst the caterpillar plot in Figure 5.3, along with the ICC value given from Equation 5.2, does suggest that there exists significant clustering amongst the repeated observations of Hb per individual, this can be validated by fitting both a null linear model and a null LME model with a random intercept and comparing them via a likelihood ratio test (LRT). An observed significant difference in these models can be attributed to the inclusion of the random intercept, therefore confirming that the observations are significantly clustered.

Similarly, the significance of a random slope can be corroborated by conducting a LRT comparing a null LME with a random intercept to a null LME with a random intercept and slope. In this case, a significant difference in the models can be attributed to the inclusion of the slope, certifying that the rate of change of the Hb levels within the CKD patients varies significantly across the individuals. The results of these LRTs are shown in Table 5.5, confirming the significance of the two random effects.

Table 5.5: Likelihood ratio tests showing the significance of the random effects for the NI dataset

	Log Likelihood	DF
Model 1: Ordinary linear model	-47034.75	2
Model 2: Random intercept model	-45141.17	3
Model 3: Random intercept and slope model	-44612.21	5
	Chi Sq.	p-value
Test 1: Model 1 and Model 2	3787.15	$< .0001$
Test 2: Model 2 and Model 3	1057.92	$< .0001$

To evaluate the effect of the observed baseline covariates on Hb, a LME model with a random intercept and slope was fitted to the NI data using the 'nlme' package within R software [183] and backward selection was used to identify the variables which had a significant effect on the individuals' Hb levels. The estimates of the significant parameters within the final model, along with their standard errors and p-values, are given in Table 5.6.

The covariance parameters of the random effects indicate a negative association between the random intercept and slope; individuals with a higher intercept typically have a more-negative slope, and thus a faster decline over time.

Table 5.6: Table showing the parameter estimates of the LME model for the NI data

	Parameter	Estimate	Standard Error	p-value
Fixed Effects	Intercept	11.065	0.120	< 0.001
	Time (Months)	-0.008	0.001	< 0.001
	Age (per 10)	0.072	0.016	< 0.001
	MCV (per 10)	-0.084	0.037	0.022
	Creatinine (per 100)	0.032	0.011	0.003
	Ferritin (per 100)	-0.043	0.006	< 0.001
Iron	Iron Hydroxide	-0.139	0.070	0.047
	Venofer	0.051	0.056	0.365
	No Iron*	-	-	-
Variance	Random Intercept	0.876	-	-
	Random Slope	0.001	-	-
	Covariance	-0.018	-	-
	Residual Error	1.378	-	-

*Baseline

This negative correlation is illustrated by the plotted random effects within Figure 5.9, where the correlation coefficient is given by $\rho = -0.763$.

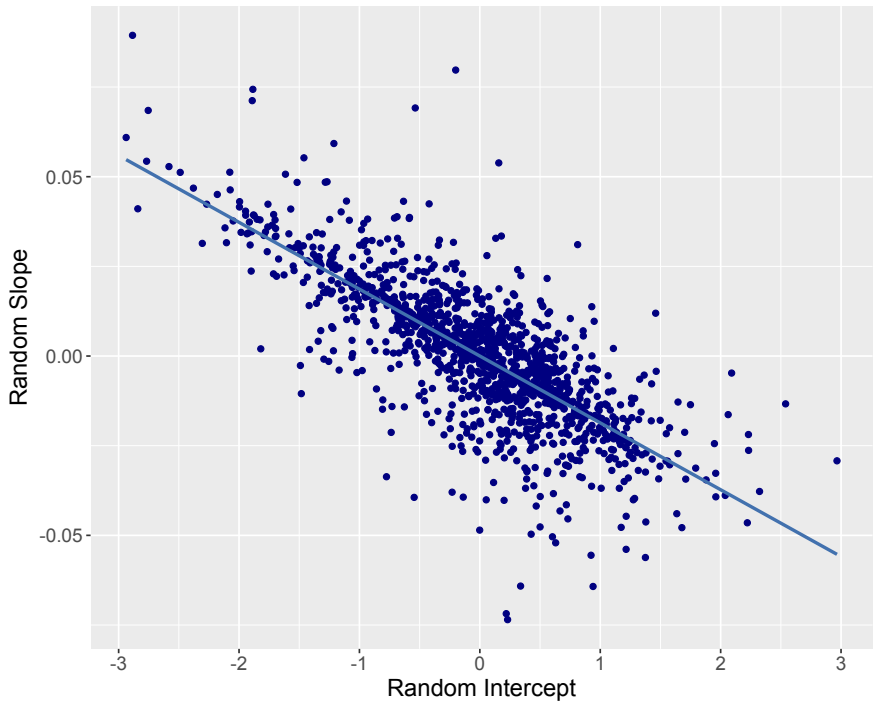


Figure 5.9: Scatter plot of the individuals' random intercepts and slopes.

To exemplify the variation that exists amongst the individuals' repeated measures trajectories, and to illustrate the fitted LME model's suitability at representing them, the estimated trajectories of nine randomly selected individuals, overlaid upon their observed Hb measures, have been plotted within Figure 5.10.

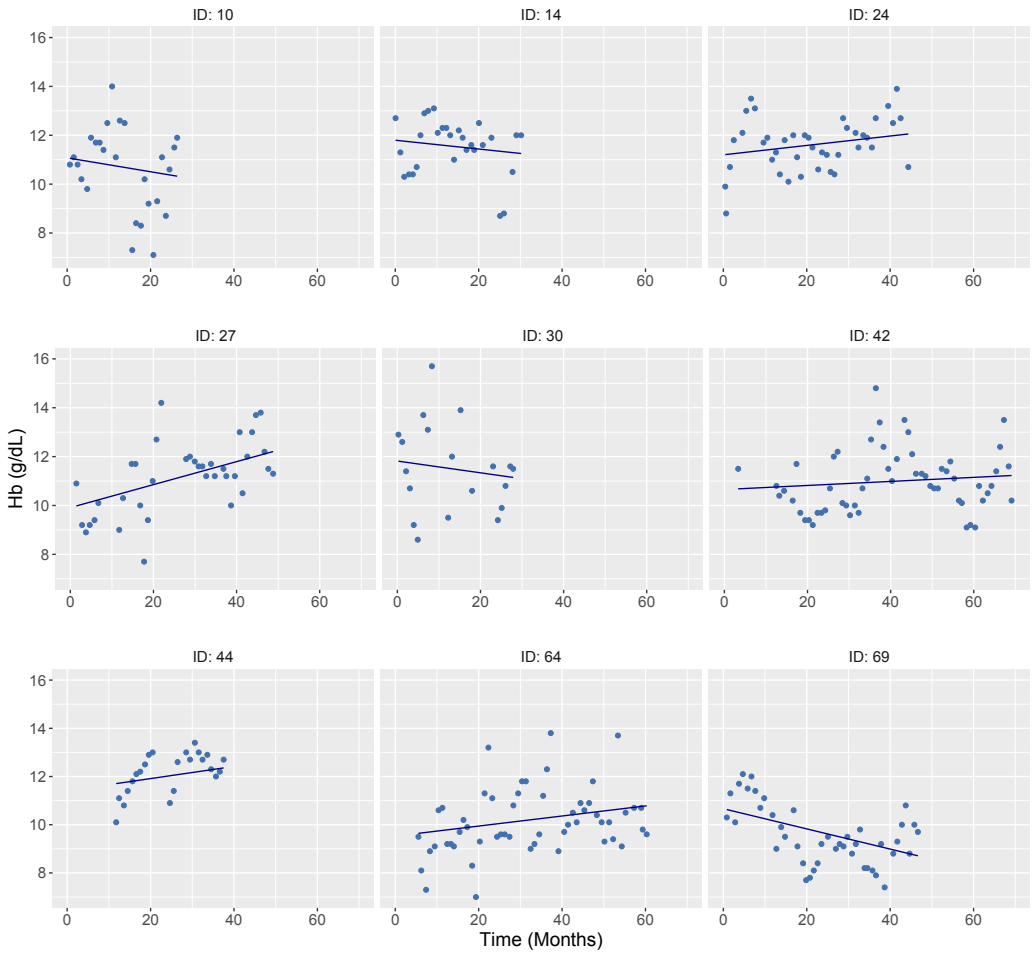


Figure 5.10: Personalised estimates of the individuals' longitudinal trajectories overlaid upon their observed values.

From these plots, it can be observed that different individuals can have very different Hb trajectories, where, for example, individual 27 has Hb levels which increase over time, compared to individual 69, whose Hb levels decrease with time. The plots also highlight different features of the data, namely that different individuals are observed for different periods of time (with different numbers of observations), and that the repeated measures per individual are similarly distributed around each individual's personalised trajectories.

5.4.1.2 Exponential, Weibull and Coxian Models for the Survival Process

Independent AFT models, assuming underlying exponential, Weibull and Coxian phase-type distributions were also fitted to the data. For fair comparison between the models, this Chapter explores only the constant effect (CE) parameterisation of the Coxian phase-type distribution, meaning all models impose the same assumption that haemoglobin has a constant effect on survival through time. The exponential and Weibull distributions were chosen for comparison due to their being the only AFT approaches currently available to represent the survival process of a joint model within the JM package in R. Baseline covariates were incorporated so as to identify those which significantly impact the survival process and to obtain initial estimates of their parameters for the joint likelihood fitting procedure (Model 1). As with the LME model, backwards selection was employed, and the final models are displayed within Tables 5.7 and 5.8, where it was found that only the individuals' baseline ages significantly impacted survival.

To investigate the impact of making naive assumptions regarding the dynamic nature of the individuals' Hb levels, two further survival models were also fitted (for each distributional assumption), incorporating an estimate of each individual's Hb level at their event time as an additional covariate within the model. Within Model 2, the individuals' Hb was estimated using LOCF and, in Model 3, the LME model previously fitted in Section 5.4.1.1 was utilised to calculate 'unbiased' estimates of the individuals' Hb at their event times, constituting a two-stage approach to the joint analysis of longitudinal and survival data. The parameter estimates from these models, adjusted for all other significant covariates, are also given in Tables 5.7 and 5.8. The exponential and Weibull models, interestingly, estimate that whilst the LOCF estimate of Hb, contaminated with error, does significantly affect survival, the predicted Hb levels from the LME model were not found to have a significant association. For the Coxian fits, however, both the LOCF and LME prediction of Hb were found to be significant.

As the shape parameter of the Weibull distribution is significant within the model, it can be inferred that the baseline distribution of the survival times is not sufficiently represented by the exponential distribution. Similarly, BIC scores show that the two-phase Coxian provides a more suitable fit than the Weibull distribution (6209.42 vs 6241.40), and that the three-phase Coxian provides a slight improvement upon the two-phase (6207.01 vs 6209.42).

Table 5.7: Parameter estimates from independent exponential and Weibull AFT survival models fitted to the CKD data, incorporating covariates (i) Model 1: Age, (ii) Model 2: Age and LOCF Hb and (iii) Model 3: Age and Predicted Hb.

		Exponential			**Weibull**		
	Parameter	**Est.**	**Std. Err.**	**p-value**	**Est.**	**Std. Err.**	**p-value**
Model 1:	Age (per 10)	-0.267	0.033	< 0.001	-0.205	0.026	< 0.001
	Intercept	6.289	0.242	< 0.001	5.740	0.196	< 0.001
	log(Shape)	-	-	-	0.259	0.033	< 0.001
Model 2:	LOCF Hb*	0.118	0.027	< 0.001	0.095	0.021	< 0.001
Model 3:	Predicted Hb*	0.045	0.056	0.445	0.023	0.042	0.582

* Adjusted for Age, Est.: Estimate, Std. Err.: Standard Error

Table 5.8: Parameter estimates from independent two, three and four phase Coxian phase-type regression models fitted to the CKD data, incorporating covariates (i) Model 1: Age, (ii) Model 2: Age and LOCF Hb and (iii) Model 3: Age and Predicted Hb.

		Two-phase Coxian			**Three-phase Coxian**			**Four-phase Coxian**		
	Parameter	**Est.**	**Std. Err.**	**p-value**	**Est.**	**Std. Err.**	**p-value**	**Est.**	**Std. Err.**	**p-value**
Model 1:	Age (per 10)	-0.217	0.023	< 0.001	-0.213	0.012	< 0.001	-0.209	0.010	< 0.001
	q_{010}	0.000	0.000	-	0.000	0.000	-	0.000	0.000	-
	q_{012}	0.032	0.005	-	0.077	0.007	-	0.102	0.008	-
	q_{020}	0.003	0.001	-	0.000	0.000	-	0.000	0.000	-
	q_{023}	-	-	-	0.071	0.006	-	0.099	0.007	-
	q_{030}	-	-	-	0.003	3×10^{-4}	-	0.000	0.000	-
	q_{034}	-	-	-	-	-	-	0.181	0.014	-
	q_{040}	-	-	-	-	-	-	0.003	3×10^{-4}	-
Model 2:	LOCF Hb*	0.103	0.018	< 0.001	0.100	0.013	< 0.001	0.101	0.010	< 0.001
Model 3:	Predicted Hb*	0.063	0.032	0.050	0.075	0.024	0.002	0.066	0.027	0.015

* Adjusted for Age, Est.: Estimate, Std. Err.: Standard Error

The inclusion of a fourth phase, however, does not significantly improve the fit (6214.48 vs 6207.01), meaning it can be concluded that the three-phase Coxian provides the best fit to the data. The estimated densities of these models are plotted within Figure 5.11, highlighting the variability which exists amongst the models. It can also be seen visually that the addition of the fourth phase within the Coxian did not significantly change the shape of the distribution, indicating that the inclusion of additional Coxian phases eventually results in conversion to a single density shape.

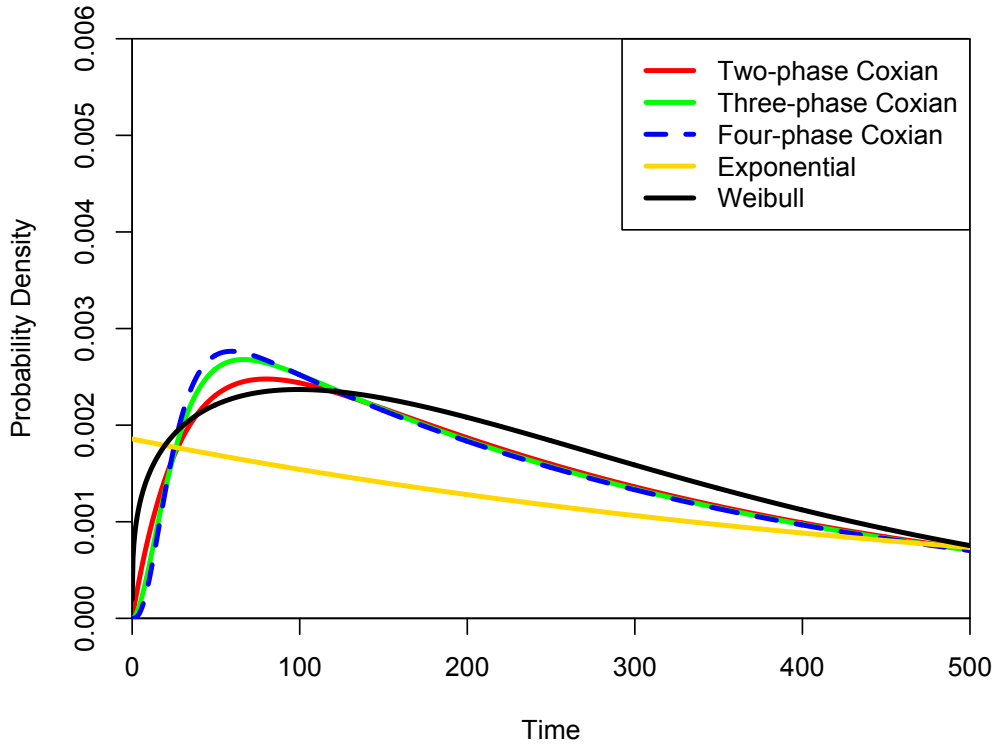


Figure 5.11: Estimated baseline probability densities from an exponential, Weibull, two-, three- and four-phase Coxian distribution.

Looking at the parameter estimates from the fitted models for the survival covariate Age, it can be observed that different specifications of the underlying distribution has a slight impact on the estimates of the survival parameter, with values ranging from -0.267 to -0.209, suggesting that a decade increase in age results in an acceleration factor between $\exp\{-\alpha\} = \exp\{0.267\} = 1.306$ and $\exp\{0.209\} = 1.232$.

5.4.2 Joint Analysis of Longitudinal and Survival Data

Within this section, the parameters of the longitudinal and survival processes are estimated simultaneously through the single joint likelihood approach. As interest lies in modelling the association between survival and the Hb levels themselves, the true longitudinal response (TLR) parameterisation of the joint likelihood is employed. As before, survival parameters are estimated according to the AFT parameterisation.

Within the JM package in R software [14], there are only two available AFT representations for the survival process: one which assumes an underlying exponential distribution and an other which assumes an underlying Weibull distribution. However, the independent survival models, fitted within Section 5.4.1.2, have already suggested that the Weibull distribution does not sufficiently represent the shape of the CKD data, meaning that the parameter estimates from a joint model which assumes the Weibull distribution are likely to be biased due to the misspecification of the survival process. Whilst alternative approaches exist for fitting PH models with atypical survival distributions, such as the piecewise constant baseline hazards model or the PH model which utilises splines to approximate the log baseline hazard, no such alternatives are currently available for the AFT model. This research aims to overcome this restriction by utilising the Coxian phase-type distribution to represent the survival process, where its ability to suitably represent any positive distribution to an arbitrary degree of accuracy should alleviate the possibility of miss-specification. As such, within this section, the two available AFT approaches within the JM package, along with the newly developed joint model which utilises the Coxian phase-type distribution, were fitted to the CKD data and their results compared.

The parameter estimates from the exponential and Weibull AFT representations are given in Table 5.10, where it can be observed that little variation is observed amongst the estimates of the longitudinal and variance parameters. The parameters associated with the survival process, however, are considerably influenced by the choice of survival model. For example, the exponential representation predicts that a one unit increase in Hb results in an acceleration factor of $\exp\{-0.658\} = 0.518$, compared to $\exp\{-0.525\} = 0.591$ from the Weibull model.

When fitting the joint model which assumes an underlying Coxian phase-type distribution, it is necessary to again determine the number of underlying phases which provide the best fit to the data, as it will not necessarily be the same number of phases as within the independent survival model. In order to do this, models were once again fitted with an increasing number of phases and the BIC was employed to compare the models and determine which provides the most suitable fit to the data. For the NI

dataset, the BIC values of the fitted models are given within Table 5.9, where it can be observed that a three-phase Coxian provides the best fit, as was the case when fitting the independent survival models.

Table 5.9: Table showing the BIC values of the fitted joint models utilising the Coxian phase-type distribution

Phase	No of Parameters	BIC
2	17	95311.20
3*	19	95310.46
4	21	95339.51

*Optimal number of phases.

The estimated parameters from this three-phase Coxian joint model are given in Table 5.10, alongside those from the exponential and Weibull AFT models. From these parameter estimates, it can again be observed that there exists little variation in the estimates of the longitudinal and variance parameters across the three models.

That is to say, these parameters are not drastically affected by the representation of the survival process. But, again, substantial differences exist in the estimates of the survival parameters, as is consistent with what has been observed previously within Simulation Study Three. Within the three-phase Coxian representation of the survival process, the acceleration factor of a one unit increase in Hb is given by $\exp\{-0.345\} = 0.708$, highlighting that the extent of this acceleration effect is overestimated by the joint models which assume an exponential or Weibull distribution.

It is interesting to note that the estimates of the Hb survival parameters from the three distributions vary significantly when compared to those of the same distribution estimated utilising a two-stage approach, displayed within Tables 5.7 and 5.8. As the two-stage approach only incorporates point estimates of the individuals' Hb levels at their death time, in comparison to the joint likelihood approach which incorporates the individuals' complete Hb trajectories, it can be concluded that the individuals' Hb medical history significantly impacts the estimates of the parameters, hence the difference in the parameter estimates.

Comparing the estimates of the effect of Hb on survival from the joint models with those from the independent survival models fitted within Section 5.4.1.2 highlights that both the LOCF Hb and predicted Hb from the LME model resulted in an underestimation of the true effect of Hb; highlighting the necessity of using joint modelling approaches when an association exists between the longitudinal and survival processes.

Table 5.10: Table showing the parameter estimates of the standard joint models available within the JM package in R applied to the NI renal data.

		Exponential			**Weibull**			**Three-phase Coxian**		
	Parameter	Est.	Std. Err.	p-value	Est.	Std. Err.	p-value	Est.	Std. Err.	p-value
Longitudinal	Intercept	11.084	0.116	< 0.001	11.078	0.118	< 0.001	11.073	0.039	< 0.001
	Time (Months)	-0.010	0.001	< 0.001	-0.010	0.001	< 0.001	-0.010	3×10^{-4}	< 0.001
	Age (per 10)	0.074	0.016	< 0.001	0.073	0.016	< 0.001	0.073	0.005	< 0.001
	MCV (per 10)	-0.093	0.035	0.008	-0.097	0.036	0.007	-0.092	0.012	< 0.001
	Creat. (per 100)	0.039	0.010	2×10^{-4}	0.038	0.011	3×10^{-4}	0.037	0.003	< 0.001
	Ferritin (per 100)	-0.043	0.006	< 0.001	-0.043	0.006	< 0.001	-0.043	0.002	< 0.001
	Iron Hydroxide	-0.139	0.069	0.044	-0.134	0.070	0.054	-0.137	0.023	< 0.001
	Venofer	0.055	0.054	0.311	0.056	0.055	0.309	0.055	0.018	0.002
	No Iron	-	-	-	-	-	-	-	-	-
Variance	Intercept	0.879	-	-	0.886	-	-	0.888	-	-
	Slope	0.001	-	-	0.001	-	-	0.001	-	-
	Covariance	-0.018	-	-	-0.018	-	-	-0.019	-	-
	Residual Error	1.376	-	-	1.376	-	-	1.376	-	-
Survival	Age (per 10)	-0.316	0.034	< 0.001	-0.246	0.027	< 0.001	-0.231	0.027	< 0.001
	Association	0.658	0.063	< 0.001	0.525	0.053	< 0.001	0.345	0.019	< 0.001
	Intercept	0.716	0.698	-	0.125	0.577	-	-	-	-
	Scale	1	-	-	1.304	-	-	-	-	-
	q_{010}	-	-	-	-	-	-	1.8×10^{-117}	6.4×10^{-61}	-
	q_{020}	-	-	-	-	-	-	8.2×10^{-32}	1.9×10^{-17}	-
	q_{030}	-	-	-	-	-	-	2.740	0.869	-
	q_{012}	-	-	-	-	-	-	0.147	0.047	-
	q_{023}	-	-	-	-	-	-	2.738	0.870	-
	Log Likelihood	-47634.57	-	-	-47608.75	-	-	-47586.83	-	-
	BIC	95377.14	-	-	95332.70	-	-	95310.46	-	-

Est.: Estimate, Std. Err.: Standard Error

A plot of the estimated baseline density corresponding to the rate parameters from the three-phase Coxian is given within Figure 5.12, and the underlying Markov process which defines the Coxian phase-type distribution uncovered by the fitted joint model is shown diagrammatically within Figure 5.13.

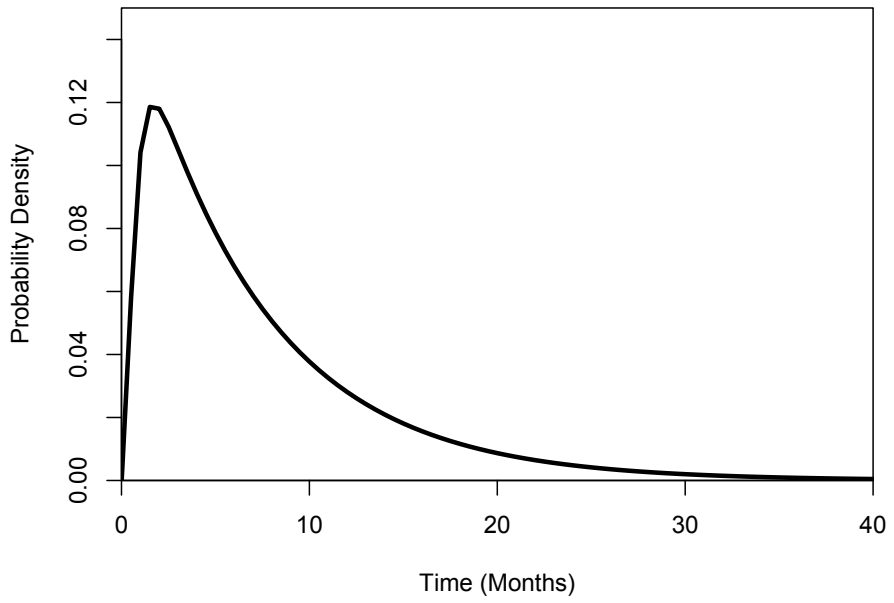


Figure 5.12: Baseline distribution of the optimal three phase Coxian for NI renal dataset.

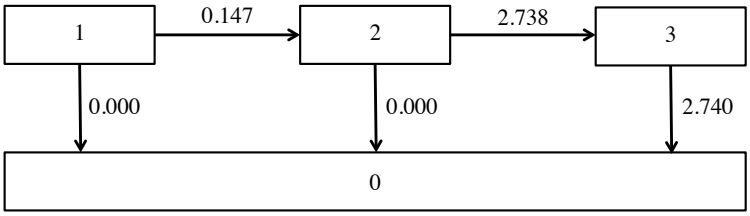


Figure 5.13: Illustration of the uncovered Markov process from the NI dataset with baseline transition intensities.

As is typical within Coxian phase-type distribution literature, it is often of interest to map these latent states onto the survival process, where in this case they can represent distinct stages of progression within end-stage renal patients. From the baseline transition intensities, it can be inferred that individuals spend the majority

of their time within the first state, with a small rate of transition into the second state of only 0.147. From this second state individuals deteriorate much faster, transitioning into the third state with rate 2.738, from which they then absorb relatively quickly, with rate 2.740. Such a rate of flow pattern suggests that the individuals within the dataset spend a prolonged period of time within relatively good health, before a rapid decline through the later diseases stages just before death. This pattern is consistent with what has been identified within previous literature [184], where, for example, Figure 5.14 shows some common health trajectories of CKD patients, which suggest that a rapid decline after an extended period of health can be typical.

As discussed previously, an advantage of employing the EM algorithm approach to fit the Coxian phase-type regression model is that, within the E-step, the expected time spent within each state is approximated for each individual, providing additional insight into how their condition evolves. For example, looking at the personalised approximations of the percentage of the individuals' survival times spent within each state, it can be identified that individuals exhibited different deterioration rates, as illustrated for three patients within Figure 5.15. Within this Figure it can be observed that, Individual 27 spent the majority of their survival time within State 1, before experiencing a rapid decline through the last two states before death, reflecting the trajectory within panel A in Figure 5.14. Conversely, Individual 42 experienced a more gradual decline towards the end of their survival time, perhaps reflecting the trajectory within panel B in Figure 5.14. Finally individual 69 spent an even more reduced proportion of time within State 1, and longer in States 2 and 3, perhaps moving toward the trajectory observed within panel C of Figure 5.14. Comparing this survival information with the individuals' Hb trajectories, plotted previously within Figure 5.10, it can be observed that Individual 27's increasing Hb levels correspond to a greater proportion of time spent in good health (i.e. State 1), whereas Individual 69's decreasing Hb trajectory corresponds to an increased time within the poorer health states.

These personalised approximations of time spent within each state can be invaluable to clinicians when determining patient intervention and making predictions regarding survival outcome; in fact, they can not only make explicit predictions on overall survival time, but also individualised predictions on the time patients will spend in each state, and thus their quality of life.

Within the CKD dataset analysed within this research, the estimated mean time spent within state one is 33.47 months, in state two is 1.58 months and state three is 1.45 months. Figure 5.16 shows three histograms of the distribution of the percentage of the individuals' survival times spent within each state, with the majority (764

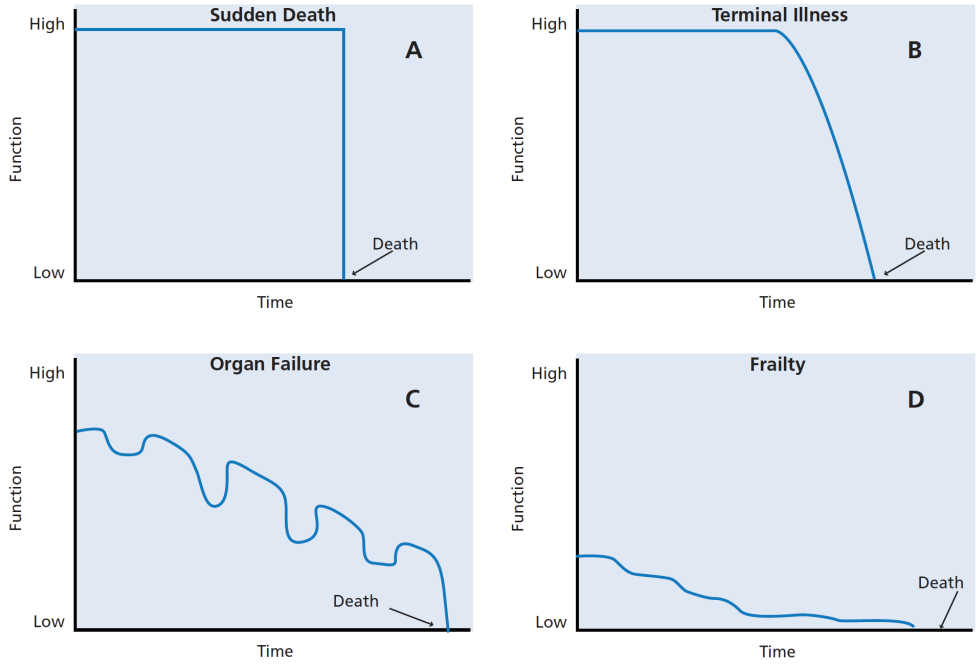


Figure 5.14: Illustration of four typical trajectories of decline within CKD patients. Source: NHS Kidney Care [184].

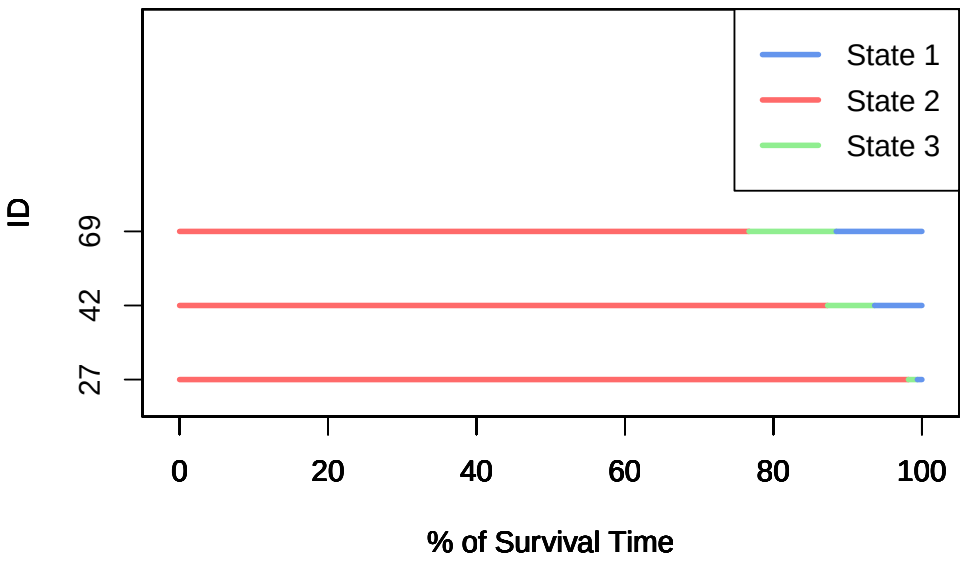


Figure 5.15: Illustration showing the estimated percentage of overall survival time spent in each state for three individuals.

individuals) spending more than 95% of their survival time in state one.

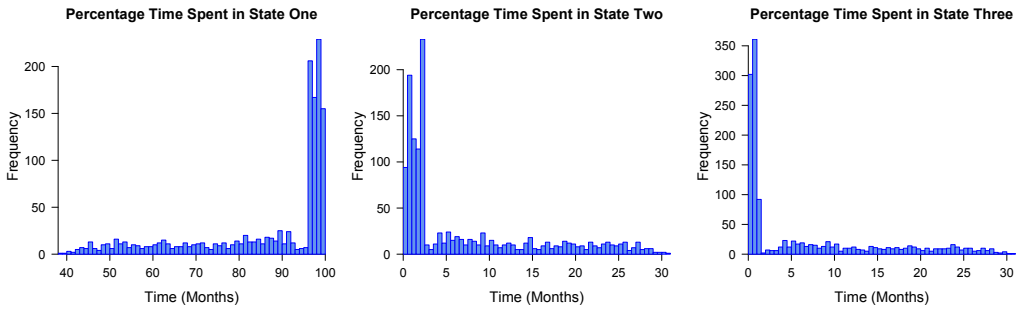


Figure 5.16: Histograms showing the percentage time spent within each of the uncovered stages of the Coxian.

5.5 Summary

Within this chapter, the newly developed joint modelling approach which utilises the Coxian phase-type regression model to represent survival was illustrated through an application to data collected from end-stage renal patients within Northern Ireland. Initially, the dataset was introduced and some preliminary analysis was presented, highlighting the possible association which exists between the haemoglobin levels of CKD patients and their survival. Subsequently, independent longitudinal and survival models were fitted to the data, before the joint likelihood approach was implemented. The newly developed approach was compared favourably to the standard joint modelling approaches available within the JM package.

Finally, some of the advantages of employing the Coxian to represent the survival process within a joint model were highlighted, such as the meaningful interpretations of the uncovered stages, and the personalised approximations regarding the times individuals will spend within each state, not available from any other survival model.

Chapter 6

Conclusion

6.1 Summary of Conclusions

This research introduces a new fully parametric joint modelling approach for the analysis of longitudinal and survival data, which utilises the Coxian phase-type regression model to represent the failure process. In doing so, this new model offers a number of advantages:

i The applicability of the Coxian phase-type regression model is extended, where it can now be employed in scenarios where interest lies in modelling the association between a survival process and a related time-varying endogenous covariate of interest. In previous literature, the Coxian phase-type regression model was restricted in terms of its scope, as it was only capable of handling time-invariant covariates, reducing its relevance within the medical field, for example, where longitudinal biomarkers are of increasing interest.

ii Within this new joint model, the Coxian phase-type distribution is highly flexible in terms of the distributional shapes that it can represent, overcoming the restrictions of alternative, fully parametric representations of the survival process, which are limited in terms of the distributional shapes they can suitably fit [7, 12, 125]. This was highlighted within Simulation Study Three in Section 4.4.3, where misspecifying the survival distribution by incorrectly employing an exponential or Weibull AFT model resulted in biased estimates of the survival parameters, which the Coxian was shown to overcome. Further, employing a fully parametric approach to represent the survival process is beneficial if interest lies in making predictions of outcome for individual patients [12], which is often a focus within medical statistics.

iii The uncovered phases of the Coxian phase-type distribution provide a new, more in-depth perception of the survival process under investigation, where inferences can be made not only on how covariates affect the survival process as a whole, but insight can also be gained into how the individuals' quality of life will evolve during their remaining survival time.

Before embarking on the development of this novel joint modelling framework, this research first focused on addressing some of the limitations of phase-type distributions, which impede their performance when representing typical survival problems. Firstly, motivated by the well documented convergence issues of phase-type distributions [39, 40, 130], a new EM algorithm approach to fitting phase-type regression models was detailed within Section 3.3. Through a simulation study, presented within Section 3.3.3, this new approach was shown to have improved accuracy in the parameter

estimates compared to previously utilised NM and QN algorithm approaches, as well as a higher rate of successful convergence, as illustrated within Figure 6.1.

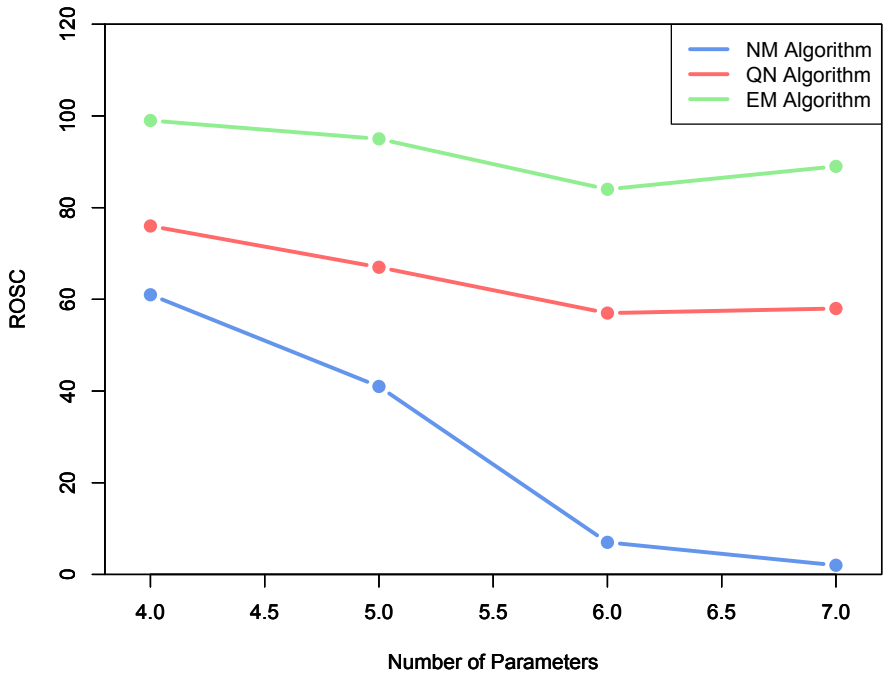


Figure 6.1: Plot showing the rate of successful convergence (ROSC) of the newly developed EM algorithm approach to fit phase-type regression models, compared with previously employed NM and QN approaches.

A further advantage of this EM algorithm approach is that, within the E-step of the model's fitting procedure, the expected time each individual spends within the different underlying states is approximated, along with the probability of visiting each state, providing insight into how individuals behave before experiencing their event of interest. This is of particular importance in the analysis of diseases in which patients experience differing qualities of life during different stages of the disease's development. For example, this is a documented issue for CKD patients, where previous research has established the importance of considering quality of life [185] and the benefits of appropriately modelling the time until end of life care is required [184].

Secondly, this research investigated relaxing a commonly made restrictive assumption regarding covariate effects within phase-type regression models, where previous approaches impose the potentially unrealistic assumption that the effect of a covariate

is constant across all transition intensities. Whilst this is done to minimise the number of covariates to be estimated by the already unstable fitting procedures, it does so at the expense of limiting the information which can be gained from the model. Detailed within Section 3.4, the increased stability of the newly developed EM approach to fitting Coxian phase-type regression models was leveraged to introduce three new parameterisations of the model which allowed (a) state-specific, (b) direction-specific and (c) transition-specific inferences to be made. Through a second simulation study, presented within Section 3.4.4, it was shown that the newly developed EM algorithm approach to these three parameterisations again outperformed the previous NM and QN approaches both in terms of the ROSC and accuracy of the parameter estimation, whilst providing detailed insight into how covariates affect different aspects of an individuals survival process.

Thirdly, in order to consider incorporating the Coxian phase-type regression model within a joint modelling framework, Section 3.5 introduced novel methodology to amend the model to allow the rate of transition between the underlying phases of the distribution to evolve over time. Where previously the underlying Markov process was considered time homogeneous, the inclusion of time-varying covariates relaxes this assumption to specify a time inhomogeneous Markov process, with baseline time-invariant hazards.

This development paved the way for the new methodological approach to fit joint models, utilising the Coxian phase-type regression model to represent the survival process. This was then detailed within Chapter 4 for the two commonly used parameterisations of the joint model, the true longitudinal response (TLR) and random effects (RE) parameterisations. A further simulation study was presented within Section 4.4.3, illustrating this new model's ability to represent more flexible distributional shapes than those joint models which assume an underlying exponential or Weibull distribution. Within the study it was also seen that bias was introduced to the estimated survival parameters when an exponential and/or Weibull distribution was assumed for data which had a more complex distributional shape, which the Coxian was shown to overcome.

This is a major advantage in the analysis of real world observational data where complicated survival densities are likely to be encountered, as illustrated by the motivating example presented within Chapter 5. Within this chapter, the new joint modelling methodology was applied to data collected from 1,340 renal patients within Northern Ireland, who received haemodialysis treatment for end-stage CKD. Through the model fitting procedure, it was determined that a three-phase Coxian provided the best fit to the data, improving upon those fits of the exponential and Weibull

distributions available within the JM package in R, through comparisons based upon the models' BIC values. From the uncovered phases of the Coxian, and the estimated rates of flow through them, it was observed that the individuals spent the majority of their survival time within the first phase of the distribution, representing the least sever stage of the disease. After transition into the second phase, individuals begin to deteriorate much more quickly, with a much higher rate of progression into the final phase, from which they were quickly absorbed. The uncovered flow pattern was found to be consistent with the anticipated behaviour of end-stage renal patients [184]. This added insight into patient behaviour, alongside the aforementioned ability to provide individualised rates of flow, as well as personalised predictions of the time spent within each state, has the potential to better inform patients and to aid clinicians in the establishment of targeted treatment plans for their patients.

6.2 Potential Further Work

Joint modelling is a rapidly-evolving area of statistical research, and as such there exists much scope for developing and improving the methodology of joint models, depending upon the desired inferences of an investigation. A particularly active aspect of this research area concerns the various procedures and techniques which can be employed to fit joint models, where the computational issues are a well discussed drawback [19]. Whilst an aspect of this research focused upon improving the issues associated with fitting the Coxian phase-type regression model through a conventional maximum likelihood approach, a potential avenue of future research could be the alternative use of Bayesian analysis in the estimation of the novel joint model which utilises the Coxian. Recent years have seen an increase in the number of Bayesian approaches employed within the field of joint modelling, beneficial due their ability to alleviate some of the computational burden of the complicated joint modelling fitting procedure [186]. Indeed, the recent publication of the 'JMBayes' package within R [20] has no doubt increased the popularity of this approach. Whilst the research presented within this thesis has explored the method of maximum likelihood to estimate the unknown parameters of the newly developed joint model, employing recent recommended approaches such as the pseudo-adaptive Gauss Hermite, it could alternatively be possible to employ Bayesian statistical approaches to compare whether an improvement in the efficiency of the model's fitting procedure could be achieved. Indeed, further investigation into the computational timings of different fitting procedures would be beneficial for both the Coxian phase-type regression model and for the newly developed joint model which utilises the Coxian.

Within this research, a LME model was employed to represent the longitudinal process, where interest was instead focused upon investigating alternative representations of the survival process. It may therefore be of interest in future work to consider previously adapted approaches explored within the literature to fit non-linear trends for the longitudinal profile, such as spline based longitudinal models [187], or a LME model with a stochastic component [188], which could be integrated alongside the Coxian within a joint modelling framework. Additionally, it could be of interest to model multiple longitudinal responses alongside a survival process, and as such it may be beneficial to consider a multivariate joint model which utilises the Coxian for the survival process.

A further extension of this work could be to investigate the commonly accepted normality assumption for the distribution of both the random effects and the residual errors within the LME model, as previous research has shown that such assumptions can be restrictive in the presence of outliers. To combat this, robust joint modelling techniques have been developed, which assume the wider-tailed t-distribution in place of normality assumptions so as to better accommodate outliers. This has been shown to improve the estimates of the longitudinal parameters [189]. As such, applying a Coxian representation of the survival process alongside a robust LME model within a joint modelling framework would reduce the bias which can occur within both processes.

Whilst the work presented within this thesis focused specifically on the application of the Coxian phase-type distribution to model survival, due to its underlying Markov process suitably representing the expected flow of individuals suffering from a chronic or degenerative condition, it could be beneficial to explore different phase-type distributions with alternative underlying Markov structures. For example, recovery from a disease could be incorporated within the model by allowing individuals to transition backwards through the underlying system, as illustrated within Figure 6.2. However, such a model, with its increased parameters, could exacerbate the identifiability issues of phase-type distributions.

As the application of the Coxian phase-type regression models to standard survival problems is a relatively novel application, it has not yet been fully explored within the literature. As such, the common adaptations to typical survival models, designed to model more complicated real-life survival situations, have not been implemented within the Coxian. For example, a competing risks scenario cannot currently be modelled using phase-type regression models. As such, there exists scope for various extensions to the Coxian phase-type regression model and thus to the novel joint model presented in this work.

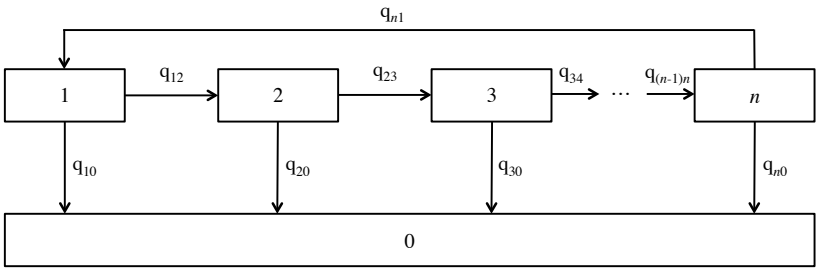


Figure 6.2: Diagram showing a Coxian phase-type distribution which allows recovery to a healthier disease state.

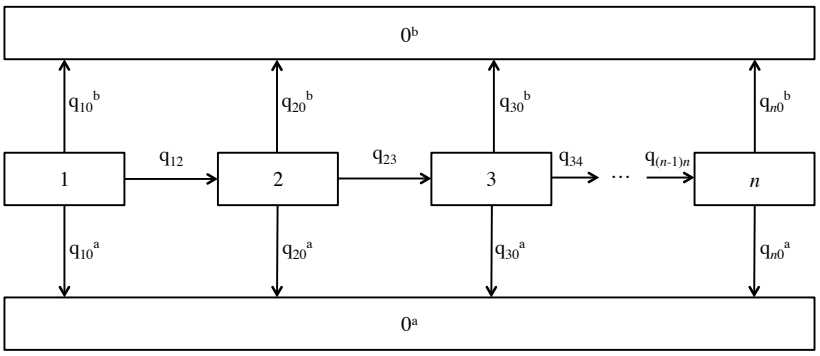


Figure 6.3: Diagram showing an underlying Markov process with two absorption states.

One possibility could be to allow individuals who experience a competing risk to transition into a second absorbing phase, as depicted within Figure 6.3. This way, all individuals who are suffering from the disease can contribute to the estimates of the rates of flow through the n transient states, representing the stages of the disease. An application of this model to a CKD scenario could be considered, where individuals often receive a kidney transplant before experiencing their event of interest. Such a model can provide insight into which event individuals are most at risk of experiencing during each stage of the disease.

Appendix A

E-Step of the Coxian Phase-type Regression Model

A.1 Expected probability of beginning the process in each state, $\mathbf{B}_{ij}$

Whilst the initialisation vector, $\mathbf{p}$, represents the overall population-average probability of beginning the process within each phase of the underlying Markov process, i.e. $\mathbf{p} = \Pr(X_0)$, personalised predictions can be obtained by considering the individuals observed event times and covariate vectors. By manipulating this conditional probability, the expected value of B_{ij} can alternatively be expressed:

$$\begin{aligned}\mathbf{E}\Big[B_{ij} \mid \tau_i\Big] &= \Pr\left(X_0 = j \mid \tau_i\right) \\ &= \frac{\Pr\left(X_0 = j\right)\Pr\left(\tau_i \mid X_0 = j\right)}{\Pr\left(\tau_i\right)},\end{aligned} \tag{A.1}$$

where

$$\Pr\left(X_0 = j\right) = p_j \tag{A.2}$$

$$\Pr\left(\tau_i \mid X_0 = j\right) = \mathbf{e_j}'\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i} \tag{A.3}$$

$$\Pr\left(\tau_i\right) = \mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i} \tag{A.4}$$

Consequently, the expected value of B_{ij} is given by:

$$\begin{aligned}\mathbf{E}\Big[B_{ij} \mid \tau_i\Big] &= \frac{p_j\mathbf{e_j}'\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}}{\mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}} \\ &= \frac{p_j d_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{p}\mathbf{d_i}(\tau_i \mid \boldsymbol{\theta}_\tau)}.\end{aligned} \tag{A.5}$$

A.2 Expected time spent within each state, $\boldsymbol{E_{ij}}$

The expected total time spent within state j of a Markov process by individual i is calculated by integrating the probability of belonging to state j with respect to time, bounded between 0 and the total time spent within the system as a whole, as shown:

$$\begin{aligned}\mathbf{E}\Big[E_{ij} \mid \tau_i\Big] &= \int_0^{\tau_i} \Pr\left(X_u = j \mid \tau_i\right) du \\ &= \int_0^{\tau_i} \frac{\Pr\left(X_u = j,\ \tau_i\right)}{\Pr\left(\tau_i\right)} \\ &= \int_0^{\tau_i} \frac{\Pr\left(\tau_i \mid X_u = j\right) \Pr\left(X_u = j\right)}{\Pr\left(\tau_i\right)} du\end{aligned} \tag{A.6}$$

where:

$$\Pr\left(X_u = j\right) = \mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}u\Big\}\mathbf{e_j} \tag{A.7}$$

$$\Pr\left(\tau_i \mid X_u = j\right) = \mathbf{e_j}' \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}(\tau_i - u)\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i} \tag{A.8}$$

and where $\Pr\left(\tau_i\right)$ is given by Equation A.4. Therefore, the expected value of E_{ij} is given by:

$$\begin{aligned}\mathbf{E}\Big[E_{ij} \mid \tau_i\Big] &= \frac{\int_0^{\tau_i} \mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}u\Big\}\mathbf{e_j}\mathbf{e_j}' \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}(\tau_i - u)\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i} du}{\mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}} \\ &= \frac{c_{ijj}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)}\end{aligned} \tag{A.9}$$

where the Runge-Kutta or Gauss-Hermite approaches can be utilised to numerically approximate this integral.

A.3 The expected probability of transitioning between each pair of states, N_{ijk}

The probability of transitioning between any pair of states j and k within a Markov model in an infinitesimal time increment, δu, conditional upon the event time is given by: $\Pr(X_u = j,\ X_{u+\delta u} = k \mid \tau_i)$. It follows that the probability of this transition occurring at some point during the time period for which the individual is known to

be in the system is given by the sum of the probabilities of the transition occurring at each infinitesimal time increment. This can be approximated by integrating the infinitesimal transition probability over the total time the individual spends within the system, as shown:

$$\begin{aligned}\mathbf{E}\Big[N_{ijk} \mid \tau_i\Big] &= \int_0^{\tau_i} \Pr\left(X_u = j,\ X_{u+\delta u} = k \mid \tau_i\right) du \\ &= \int_0^{\tau_i} \frac{\Pr\left(X_u = j,\ X_{u+\delta u} = k,\ \tau_i\right)}{\Pr\left(\tau_i\right)} du \\ &= \int_0^{\tau_i} \frac{\Pr\left(X_u = j\right) \Pr\left(X_{u+\delta u} = k \mid X_u = j\right) \Pr\left(\tau_i \mid X_u = j,\ X_{u+\delta u} = k\right)}{\Pr\left(\tau_i\right)}\end{aligned} \tag{A.10}$$

where:

$$\Pr\left(X_u = j\right) = \mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}u\Big\}\mathbf{e_j} \tag{A.11}$$

$$\Pr\left(X_{u+\delta u} = k \mid X_u = j\right) = q_{0jk} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\} \tag{A.12}$$

$$\begin{aligned}\Pr\left(\tau_i \mid X_u = j,\ X_{u+\delta u} = k\right) &\equiv \Pr\left(\tau_i \mid X_{u+\delta u} = k\right) \\ &= \mathbf{e_k}' \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}(\tau_i - u - \delta u)\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}\end{aligned} \tag{A.13}$$

and where $\Pr\left(\tau_i\right)$ is given by Equation A.4. This means that the expected value of N_{ijk} is given by:

$$\begin{aligned}\mathbf{E}\Big[N_{ijk} \mid \tau_i\Big] &= \frac{1}{\mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}} \\ &\quad \times \left(\int_0^{\tau_i} \mathbf{p} \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}u\Big\}\mathbf{e_j} q_{0jk} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\mathbf{e_k}' \right. \\ &\qquad \left. \times \exp\Big\{\mathbf{T} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}(\tau_i - u)\Big\}\Big(\mathbf{t} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i} du \right) \\ &= \frac{c_{ijk}(\tau_i \mid \boldsymbol{\theta}_\tau) q_{0jk} \exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \qquad \text{as } \delta u \to 0.\end{aligned} \tag{A.14}$$

A.4 Expected probability of absorbing from each state, N_{ij0}

The expected probability of absorbing from state j, i.e the expected probability of belonging to state j at the infinitesimal time increment before absorption, $\tau_i - \delta u$, can be expressed as:

$$\begin{aligned}\mathbf{E}\Big[N_{ij0} \mid \tau_i\Big] &= \Pr\left(X_{\tau_i-\delta u} = j \mid \tau_i\right) \\ &= \frac{\Pr\left(X_{\tau_i-\delta u} = j,\ \tau_i\right)}{\Pr\left(\tau_i\right)} \\ &= \frac{\Pr\left(X_{\tau_i-\delta u} = j\right)\Pr\left(\tau_i \mid X_{\tau_i-\delta u} = j\right)}{\Pr\left(\tau_i\right)}\end{aligned} \tag{A.15}$$

where:

$$\Pr\left(X_{\tau_i-\delta u} = j\right) = \mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}(\tau_i - \delta u)\Big\}\mathbf{e_j} \tag{A.16}$$

$$\Pr\left(\tau_i \mid X_{\tau_i-\delta u} = j\right) = \mathbf{e_j}'\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\delta u\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i} \tag{A.17}$$

and where $\Pr\left(\tau_i\right)$ is given by Equation A.4. Thus, the expected value of N_{ij0} is given by:

$$\begin{aligned}\mathbf{E}\big[N_{ij0} \mid \tau_i\big] &= \frac{\mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}(\tau_i - \delta u)\Big\}\mathbf{e_j}\mathbf{e_j}'\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\delta u\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}}{\mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}} \\ &= \frac{\mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\mathbf{e_j}q_{0j0}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}}{\mathbf{p}\exp\Big\{\mathbf{T}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\tau_i\Big\}\Big(\mathbf{t}\exp\{-\boldsymbol{w}_i\boldsymbol{\gamma}\}\Big)^{\delta_i}} \\ &= \frac{a_{ij}(\tau_i \mid \boldsymbol{\theta}_\tau)}{\mathbf{pd_i}(\tau_i \mid \boldsymbol{\theta}_\tau)} \qquad \text{as } \delta u \to 0\end{aligned} \tag{A.18}$$

Bibliography

[1] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[3] C. L. Faucett and D. C. Thomas. Simultaneously modelling censored survival data and repeatidly measured covariates: A gibbs sampling approach. *Statistics in Medicine*, 15:1663–1685, 1996.

[4] M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997.

[5] A. A. Tsiatis, U. Dafni, V. De Gruttola, K. J. Propert, R. L. Strawderman, and M. Wulfsohn. The relationship of CD4 counts over time to survival in patients with AIDS: Is CD4 a good surrogate marker? In N. P. Jewell, K. Dietz, and V. T. Farewell, editors, *AIDS Epidemiology*, pages 256–274. Birkhauser Boston, 1992.

[6] S. Self and Y. Pawitan. Modeling a marker of disease progression and onset of disease. In N. P. Jewell, K. Dietz, and V. T. Farewell, editors, *AIDS Epidemiology*, pages 231–255. Birkhauser Boston, 1992.

[7] M. J. Crowther, K. R. Abrams, and P. C. Lambert. Flexible parametric joint modelling of longitudinal and survival data. *Statistics in Medicine*, 31(30):4456–4471, 2012.

[8] V. De Gruttola and X. M. Tu. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4):1003–1014, 1994.

[9] R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

[10] Y. K. Tseng, F. Hsieh, and J. L. Wang. Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, 92(3):587–603, 2005.

[11] F. Hsieh, Y. K. Tseng, and J. L. Wang. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(4):1037–1043, 2006.

[12] A. L. Gould, M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, 34(14):2181–2195, March 2014.

[13] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling*. Springer-Verlag, Berlin, Heidelberg, 1995.

[14] D. Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[15] Thomson Reuters Web of Knowledge. http://apps.webofknowledge.com., September 2013.

[16] D. Rizopoulos, G. Verbeke, and E. Lesaffre. Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654, 2009.

[17] D. Rizopoulos. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis*, 56(3):491 – 501, 2012.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[19] D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, 2012.

[20] D. Rizopoulos. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–45, 2016.

[21] P. Philipson, I. Sousa, P. J. Diggle, P. Williamson, R. Kolamunnage-Dona, R. Henderson, and L. Hickey. *joineR: Joint Modelling of Repeated Measurements and Time-to-Event Data*, 2018. R package version 1.2.3.

[22] M. F. Neuts. *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. Algorithmic Approach. Dover Publications, 1981.

[23] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineers Journal*, 10:189–197, 1917.

[24] T. L. Saaty. *Elements of queueing theory: with applications*. McGraw-Hill, 1961.

[25] A. H. Marshall and B. Shaw. Fitting coxian phase-type distributions to patient length of stay. In *Paper presented at Conference on Applied Statistics in Ireland*, pages 39–41, 2003.

[26] A. H. Marshall, C. Vasilakis, and E. El-Darzi. Length of stay-based patient flow models: Recent developments and future directions. *Health Care Management Science*, 8(3):213–220, 2005.

[27] M. Zenga, A. H. Marshall, and G. Sabrina. Modelling students' length of stay at university using Coxian phase-type distributions. *International Journal of Statistics and Probability*, 2(1):73, 2013.

[28] O. O. Aalen. Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, 22(4):447–463, 1995.

[29] C. Donnelly, L. M. McFetridge, A. H. Marshall, and H. J. Mitchell. A two-stage approach to the joint analysis of longitudinal and survival data utilising the coxian phase-type distribution. *Statistical Methods in Medical Research*, 2017.

[30] X. Tang, Z. Luo, and J.C. Gardiner. Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. *Statistics in Medicine*, 31(14):1502–1516, 2012.

[31] M. Olsson. Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, 23(4):pp. 443–460, 1996.

[32] Louis Aslett. *PhaseType: Inference for Phase-type Distributions*, 2012. R package version 0.1.3.

[33] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996.

[34] UK Renal Registery. https://www.renalreg.org/, September 2018.

[35] M. Kerr, B. Bray, J. Medcalf, D. J. O'Donoghue, and B. Matthews. Estimating the financial cost of chronic kidney disease to the NHS in England. *Nephrology Dialysis Transplantation*, 27((Suppl 3)):73–80, 2012.

[36] J. Neugarten and L. Golestaneh. Gender and the prevalence and progression of renal disease. *Advances in Chronic Kidney Disease*, 20(5):390 – 395, 2013. Women in Nephrology.

[37] World Kidney Day. www.worldkidneyday.org, Aug 2018.

[38] M. Tonelli and M. Riella. Chronic kidney disease and the aging population. *Brazilian Journal of Nephrology*, 36:1 – 5, 03 2014.

[39] C. A. O'Cinneide. On non-uniqueness of representations of phase-type distributions. *Communications in Statistics. Stochastic Models*, 5(2):247–259, 1989.

[40] A. Lang and J. Arthur. Parameter approximation for phase-type distributions. In *Lecture Notes in Pure and Applied Mathematics*, pages 151–206–. CRC Press, September 1996.

[41] P. J. Diggle, K. Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 1994.

[42] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley, 2012.

[43] H. Goldstein. Age, period and cohort effects - a confounded confusion. *Journal of Applied Statistics*, 6(1):19–24, 1979.

[44] E. J. Caruana, M. Roman, J. Hernandez-Sanchez, and P. Solli. Longitudinal studies. *Journal of Thoracic Disease*, 7(11):537–540, Oct 2015.

[45] Lekisha Edwards. Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatric pulmonology*, 30 4:330–44, 2000.

[46] E. Demidenko. *Mixed Models: Theory and Applications with R*. Wiley Series in Probability and Statistics. Wiley, 2013.

[47] R.A. Fisher. *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd, 1928.

[48] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[49] D. Hedeker and R. D. Gibbons. *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2006.

[50] A. Gevins, M. E. Smith, L. McEvoy, and D. Yu. High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, 7(4):374–385, 1997.

[51] H. E. Woodman, R. E. Evans, E. H. Callow, and J. Wishart. The nutrition of the bacon pig. i. the influence of high levels of protein intake on growth, conformation and quality in the bacon pig. *The Journal of Agricultural Science*, 26(4):546 – 619, 1936.

[52] J. Wishart. Growth-rate determinations in nutrition studies with bacon pig, and their analysis. *Biometrika*, 30(1-2):16–28, 1938.

[53] RA Fisher. *The Design of Experiments*. Hafner Pub. Co., 1935.

[54] H. Scheffe. *The Analysis of Variance*. A Wiley publication in mathematical statistics. Wiley, 1959.

[55] G. M. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis, 2008.

[56] S. W. Greenhouse and S. Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959.

[57] C. R. Henderson. Selection index and expected genetic advance. *Statistical genetics and plant breeding*, 982:141–163, 1963.

[58] R. Gueorguieva and J. H. Krystal. Move over ANOVA: Progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry. *Archives of General Psychiatry*, 61(3):310–317, 2004.

[59] S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494, 1932.

[60] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 1984.

[61] G. E. P. Box. Problems in the analysis of growth and wear curves. *Biometrics*, 6(4):362–389, 1950.

[62] R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3/4):313–326, 1964.

[63] C. R. Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4):447–458, 1965.

[64] J. E. Grizzle and D. M. Allen. Analysis of growth and dose response curves. *Biometrics*, 25(2):357–381, 1969.

[65] G. Reinsel. Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, 77(377):190–195, 1982.

[66] R. D. Gibbons, D. Hedeker, and S. DuToit. Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, 6(1):79–107, 2010.

[67] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of The American Statistical Association*, 72(358):320–338, 1977.

[68] C. H. Morrell, L. J. Brant, S. Sheng, and E. J. Metter. Screening for prostate cancer using multivariate mixed-effects models. *Journal of Applied Statistics*, 39(6):1151–1175, 2012.

[69] Y. Huang and G. Dagne. Comparison of mixed-effects models for skew-normal responses with an application to aids data: A bayesian approach. *Communications in Statistics: Simulation & Computation*, 42(6):1268 – 1287, 2013.

[70] V. M. Muggeo, D. C. Atkins, R. J. Gallop, and S. Dimidjian. Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling: An International Journal*, 14(4):293–313, 2014.

[71] J. Chao, L. Yang, H. Xu, Q. Yu, L. Jiang, and M. Zong. The effect of integrated health management model on the health of older adults with diabetes in a randomized controlled trial. *Archives of Gerontology and Geriatrics*, 60:82 – 88, 2015.

[72] N Bennett. *Teaching styles and pupil progress*. Open Books, 1976.

[73] J. Gray and D. Satterly. A chapter of errors: Teaching styles and pupil progress in retrospect. *Educational Research*, 19(1):45–56, 1976.

[74] M. Aitkin and N. Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, 149(1):1–43, 1986.

[75] N. M. Laird. Computation of variance components using the em algorithm. *Journal of Statistical Computation and Simulation*, 14(3-4):295–303, 1982.

[76] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[77] A. P. Dempster, D. B. Rubin, and R. K. Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374):341–353, 1981.

[78] N. M. Laird, N. Lange, and D. Stram. Maximum likelihood computations with repeated measures: Application of the em algorithm. *Journal of the American Statistical Association*, 82(397):97–105, 1987.

[79] P. A. V. B. Swamy. *Statistical inference in random coefficient regression models.* Lecture notes in operations research and mathematical systems. Springer-Verlag, 1971.

[80] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[81] D. A. Harville. Extension of the gauss-markov theorem to include the estimation of random effects. *The Annals of Statistics*, 4(2):384–395, 1976.

[82] R. Stiratelli, N. M. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971, 1984.

[83] E. L. Korn and A. S. Whittemore. Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, 35(4):795–802, 1979.

[84] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687, 1990.

[85] R. D. Gibbons and D. Hedeker. Applications of mixed-effects models in biostatistics. *The Indian Journal of Statistics, Series B*, 62(1):70–103, 2000.

[86] R. D. Gibbons and D. Hedeker. Random effects probit and logistic regression models for three-level data. *Biometrics*, 53(4):1527–1537, 1997.

[87] D. Collett. *Modelling Survival Data in Medical Research, Second Edition.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003.

[88] R. Singh and K. Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, 2(4):145–148, 2011.

[89] P. Shamsi. On-line survival analysis of power electronic converters using step noise-cox processes. *CoRR*, abs/1507.06000, 2015.

[90] A. Kassani, M. Niazi, J. Hassanzadeh, and R. Menati. Survival analysis of drug abuse relapse in addiction treatment centers. *International Journal of High Risk Behaviors & Addiction*, 4(3), 2015.

[91] F. Siannis, J. Copas, and G. Lu. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, 6(1):77–91, 2005.

[92] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer, 2003.

[93] P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[94] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 2002.

[95] R. L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

[96] A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834, 2004.

[97] L. McCrink, A. H. Marshall, K. Cairns, D. Fogarty, and A. Casula. Joint modelling of longitudinal and survival data: A comparison of joint and independent models. In *Proc 58th World Statistical Congress*, 2011.

[98] L. Wu, W. Liu, G. Y. Yi, and Y. Huang. Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012.

[99] D. Follmann and M. Wu. An approximate generalized linear model with random effects for informative missing data. *Biometrics*, 51(1):151–168, 1995.

[100] A. A. Tsiatis, V. De Gruttola, and M. S. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37, 1995.

[101] Y. Pawitan and S. Self. Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, 88:719–726, 09 1993.

[102] V. De Gruttola and X. M. Tu. Modeling the relationship between progression of CD4-lymphocyte count and survival time. In N. P. Jewell, K. Dietz, and V. T. Farewell, editors, *AIDS Epidemiology*, pages 275–296. Birkhauser Boston, 1992.

[103] U. G. Dafni and A. A. Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54(4):1445–1462, 1998.

[104] P. Bycott and J. Taylor. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine*, 17(18):2061–2077, 1998.

[105] M. C. Wu and R. J. Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188, 1988.

[106] M. J. Sweeting and S. G. Thompson. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763, 2011.

[107] J. Hogan and N. Laird. Model-based approached to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16(3):259–272, 1997.

[108] L. M. McCrink, A. H. Marshall, and K. J. Cairns. Advances in joint modelling: A review of recent developments with application to the survival of end stage renal disease patients. *International Statistical Review*, 81(2):249–269, 2013.

[109] R. J. Glynn, N. M. Laird, and D. B. Rubin. *Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse*, pages 115–142. Springer New York, New York, NY, 1986.

[110] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, 2002.

[111] R. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[112] R. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

[113] R. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

[114] I. Sousa. A review on joint modelling of longitudinal measurements and time-to-event. *Revstat*, 9(1):57–81, 2011.

[115] L. Wu. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies. *Journal of the American Statistical Association*, 97(460):955–964, 2002.

[116] K. Salah, A.B Mohd, N. Ibrahim, and K. Haron. A stochastic joint model for longitudinal and survival data with cure patients. *International Journal of Tomography & Statistics*, 11, 10 2009.

[117] E. R. Brown, J. G. Ibrahim, and V. DeGruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73, 2005.

[118] A. S. Whittemore and J. B. Keller. Survival estimation using splines. *Biometrics*, 42(3):495–506, 1986.

[119] P. S. Rosenberg. Hazard function estimation using b-splines. *Biometrics*, 51(3):874–887, 1995.

[120] J. Herndon and F. Harrell. The restricted cubic spline hazard model. *Communications in Statistics - Theory and Methods*, 19(2):639–663, 1990.

[121] S. Asmussen. *Applied Probability and Queues*. Applications of mathematics : stochastic modelling and applied probability. Springer, 2003.

[122] X. Lin, J. Taylor, and W. Ye. A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and its Interface*, 1(1):33–45, 2008.

[123] W. H. Press. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007.

[124] M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Probability: Pure and Applied. Taylor & Francis, 1989.

[125] M. J. Faddy. Phase-type distributions for failure times. *Mathematical and computer modelling*, 22(10-12):63–70, 1995.

[126] K. Smaili, T. Kadri, and S. Kadry. Hypoexponential distribution with different parameters. *Applied Mathematics*, 4(4):pp. 624–631, 2013. doi: 10.4236/am.2013.44087.

[127] D.R. Cox and H.D. Miller. *The Theory of Stochastic Processes*. Science paperbacks. Taylor & Francis, 1977.

[128] C. H. Jackson. Multi-state models for panel data: the msm package for r. *Journal of Statistical Software*, 38(8):1–29, 2011.

[129] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

[130] M. J. Faddy. On inferring the number of phases in a Coxian phase-type distribution. *Communications in Statistics. Stochastic Models*, 14(1-2):407–417, 1998.

[131] M. J. Faddy and S. I. McClean. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15(4):311–317, 1999.

[132] M. J. Faddy and S. I. McClean. Markov chain modelling for geriatric patient care. *Methods Archive*, 44(3):369–373, 2005.

[133] A. H. Marshall and S. I. McClean. Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Science*, 7(4):285–289, 2004.

[134] M. F. Neuts, R. Perez-Ocon, and I. Torres-Castro. Repairable models with operating and repair times governed by phase type distributions. *Advances in Applied Probability*, 32(2):468–479, 2000.

[135] S. Ahn, J. H. Kim, and V. Ramaswami. A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics*, 51(1):43 – 52, 2012.

[136] A. Kolmogoroff. Uber die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931.

[137] H. Xie. *Modelling Issues in Institutional Long-term Care: Placement, Survival and Cost*. PhD thesis, University of Westminster, 2004.

[138] A. H. Marshall and M. Zenga. Experimenting with the Coxian phase-type distribution to uncover suitable fits. *Methodology and Computing in Applied Probability*, 14(1):71–86, 2010.

[139] M. Fackrell. Modelling healthcare systems with phase-type distributions. *Health care management science*, 12(1):11, 2009.

[140] A. H. Marshall and S. I. McClean. Conditional phase-type distributions for modelling patient length of stay in hospital. *International Transactions in Operational Research*, 10(6):565–576, 2003.

[141] M. A. Johnson. Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and Erlang distributions. *ORSA Journal on Computing*, 5(1):69–83, 1993.

[142] A. Riska, V. Diev, and E. Smirni. Efficient fitting of long-tailed data sets into hyperexponential distributions. In *Global Telecommunications Conference, 2002. GLOBECOM 02. IEEE*, volume 3, pages 2513–2517 vol.3, Nov 2002.

[143] M. J. Faddy. A structured compartmental model for drug kinetics. *Biometrics*, 49(1):243–248, 1993.

[144] A. H. Marshall and M. Zenga. Recent developments in fitting Coxian phase-type distributions in healthcare. In *In Applied Stochastic Models and Data Analysis : the XIII International Conference.*, 2009.

[145] K. Payne, A. H. Marshall, and K. J. Cairns. Investigating the efficiency of fitting coxian phase-type distributions to health care data. *IMA Journal of Management Mathematics*, 23(2):133–145, 2012.

[146] A. H. Marshall and M. Zenga. Simulating Coxian phase-type distributions for patient survival. *International Transactions in Operational Research*, 16(2):213–226, 2009.

[147] M. J. Faddy, N. Graves, and A. Pettitt. Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314, 2009.

[148] C. A. McGrory, A. N. Pettitt, and M. J. Faddy. A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis*, 53(12):4311 – 4321, 2009.

[149] G. Marshall and R. H. Jones. Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14(18):1975–1983, 1995.

[150] M. C. Ausin, M. P. Wiper, and R. E. Lillo. Bayesian prediction of the transient behaviour and busy period in short- and long-tailed queueing systems. *Computational Statistics & Data Analysis*, 52(3):1615 – 1635, 2008.

[151] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* Wiley series in probability and statistics. Wiley-Interscience, 2008.

[152] S. Asmussen. Phase-type distributions and related point processes: Fitting and recent advances. In *Lecture Notes in Pure and Applied Mathematics*, pages 137–149. CRC Press, 1996.

[153] C. Dutang, V. Goulet, and M. Pigeon. actuar: An r package for actuarial science. *Journal of Statistical Software, Articles*, 25(7):1–37, 2008.

[154] H. Okamura and T. Dohi. Building phase-type software reliability models. In *2006 17th International Symposium on Software Reliability Engineering*, pages 289–298, Nov 2006.

[155] Andreas Lang and J L. Arthur. Parameter approximation for phase-type distributions. 09 1996.

[156] M. Kitahata, S. J. Gange, A. G. Abraham, B. Merriman, M. S. Saag, A. C. Justice, R. S. Hogg, S. G. Deeks, J. J. Eron, J. T. Brooks, S. B. Rourke, M. J. Gill, R. J. Bosch, J. N. Martin, M. B. Klein, L. P. Jacobson, B. Rodriguez, T. R. Sterling, G. D. Kirk, S. Napravnik, A. R. Rachlis, L. M. Calzavara, M. A. Horberg, M. J. Silverberg, K. A. Gebo, J. J. Goedert, C. A. Benson, A. C. Collier, S. E. Van Rompaey, H. M. Crane, R. G. McKaig, B. Lau, A. M. Freeman, and R. D. Moore. Effect of early versus deferred antiretroviral therapy for hiv on survival. *New England Journal of Medicine*, 360(18):1815–1826, 2009. PMID: 19339714.

[157] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3):245 – 279, 1998.

[158] A. Riska, V. Diev, and E. Smirni. Efficient fitting of long-tailed data sets into phase-type distributions. *SIGMETRICS Perform. Eval. Rev.*, 30(3):6–8, December 2002.

[159] A. Riska, V. Diev, and E. Smirni. An em-based technique for approximating long-tailed data sets with ph distributions. *Performance Evaluation*, 55(1):147 – 164, 2004. Internet Performance Symposium (IPS 2002).

[160] M. Bladt. A review on phase-type distributions and their use in risk theory. *ASTIN Bulletin*, 35(1):145–161, 2005.

[161] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. A comparison of joint models for longitudinal and competing risks data, with application to an epilepsy drug randomized controlled trial. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 0(0), 2018.

[162] B. He and S. Luo. Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinson's disease. *Statistical methods in medical research*, 25(4):1346–1358, 2016.

[163] M. J. J Crowther, T. M.-L. Andersson, P. C. Lambert, K. R. Abrams, and K. Humphreys. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in medicine*, 35(7):1193–1209, 2016.

[164] National Kidney Foundation. www.kidney.org, Aug 2018.

[165] National Institute of Diabetes, Digestive, and Kidney Diseases. www.niddk.nih.gov, Aug 2018.

[166] The Renal Association. www.renal.org, Aug 2018.

[167] UNC Kidney Centre. www.unckidneycentre.org, Aug 2018.

[168] H. Gray. *Gray's Anatomy: The Classic First Edition*. Longmeadow Press, 1994.

[169] National Health Service. www.nhs.uk/conditions/kidney-disease, Aug 2018.

[170] M. Kerr. Chronic kidney disease in England: The human and financial cost. *NHS Kidney Care*, 2012.

[171] L. A. Stevens, J. Coresh, T. Greene, and A. S. Levey. Assessing kidney function — measured and estimated glomerular filtration rate. *New England Journal of Medicine*, 354(23):2473–2483, 2006. PMID: 16760447.

[172] A. S. Levey, J. Coresh, T. Greene, J. Marsh, L. A. Stevens, J. W. Kusek, and F. Van Lente. Expressing the modification of diet in renal disease study equation for estimating glomerular filtration rate with standardized serum creatinine values. *Clinical Chemistry*, 53(4):766–772, 2007.

[173] R. A. McPherson and M. R. Pincus. *Henry's Clinical Diagnosis and Management by Laboratory Methods E-Book*. Elsevier Health Sciences, 2011.

[174] J. L. Babitt and H. Y. Lin. Mechanisms of anemia in CKD. *Journal of the American Society of Nephrology*, 23(10):1631–1634, 2012.

[175] G. Tsagalis. Renal anemia: a nephrologist's view. *Hippokratia*, 15(Suppl 1):39–43, January 2011.

[176] Kidney Research UK. www.kidneyresearchuk.org, Aug 2018.

[177] L. A. Stevens, G. Viswanathan, and D. E. Weiner. Chronic kidney disease and end-stage renal disease in the elderly population: Current prevalence, future projections, and clinical significance. *Advances in Chronic Kidney Disease*, 17(4):293 – 301, 2010. Aging and Chronic Kidney Disease.

[178] I. Goldberg and l. Krause. The role of gender in chronic kidney disease. *EMJ*, 1(2):58–64, 2016.

[179] K. Kalantar-Zadeh, K. Kalantar-Zadeh, and G. H. Lee. The fascinating but deceptive ferritin: To measure it or not to measure it in chronic kidney disease? *Clinical Journal of the American Society of Nephrology*, 1(Supplement 1):S9–S18, 2006.

[180] S. Gowda, P. B. Desai, S. S. Kulkarni, V. V. Hull, A. A. K. Math, and S. N. Vernekar. Markers of renal function tests. *North American Journal of Medical Sciences*, 2(4):170–173, April 2010.

[181] Z. Jing, Y. Wei-jie, Z. Nan, Z. Yi, and W. Ling. Hemoglobin targets for chronic kidney disease patients with anemia: A systematic review and meta-analysis. *PLoS ONE*, 7(8), 2012.

[182] M. Z. Molnar, U. Mehrotra, R. Mehrotra, C. P. Kovesdy, and K. Kalantar-Zadeh. Association of hemoglobin and survival in peritoneal dialysis patients. *Clinical Journal of the American Society of Nephrology*, 6(8):1973–1981, 2011.

[183] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2018. R package version 3.1-137.

[184] NHS Kidney Care. End of life care in advanced kidney disease: A framework for implementation. *National End of Life Care Progamme*, 2015.

[185] M. C. Cruz, C. Andrade, M. Urrutia, S. Draibe, L. A. Nogueira-Martins, and R. de Castro Cintra Sesso. Quality of life in patients with chronic kidney disease. *Clinics*, 66(6):991–995, March 2011.

[186] A. H. Dessiso and A. T. Goshu. Bayesian joint modelling of longitudinal and survival data of hiv/aids patients: A case study at bale robe general hospital, ethiopia. *American Journal of Theoretical and Applied Statistics*, 6(4):182–190, 2017.

[187] P. T. T. Huong, D. Nur, and A. Branford. Penalized spline joint models for longitudinal and time-to-event data. *Communications in Statistics - Theory and Methods*, 46(20):10294–10314, 2017.

[188] C. Mbogning, K. Bleakley, and M. Lavielle. Joint modelling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the stochastic approximation expectation-maximisation algorithm. *Journal of Statistical Computation and Simulation*, 85(8):1512–1528, 2015.

[189] N. Li, R. M. Elashoff, and G. Li. Robust joint modeling of longitudinal measurements and competing risks failure time data. *Biometrical journal. Biometrische Zeitschrift*, 51(1):19–30, February 2009.